



Estimation par sélection de modèle en régression hétéroscédastique

Xavier Gendre

► To cite this version:

Xavier Gendre. Estimation par sélection de modèle en régression hétéroscédastique. Mathématiques [math]. Université Nice Sophia Antipolis, 2009. Français. NNT : . tel-00397608

HAL Id: tel-00397608

<https://theses.hal.science/tel-00397608>

Submitted on 22 Jun 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE NICE SOPHIA ANTIPOLIS – UFR Sciences

École Doctorale Sciences Fondamentales et Appliquées

THÈSE

pour obtenir le titre de
Docteur en Sciences
Spécialité : MATHÉMATIQUES

présentée et soutenue par
Xavier GENDRE

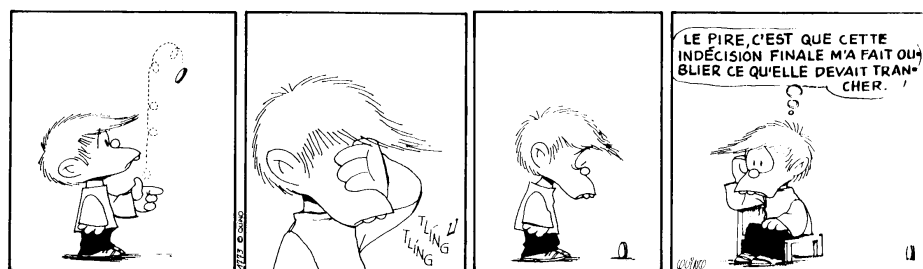
Estimation par sélection de modèle en régression hétéroscédastique

Thèse dirigée par **Yannick BARAUD**
soutenue le 15 juin 2009

Membres du jury :

M.	Yannick	BARAUD	Directeur de thèse
M.	Lucien	BIRGÉ	Président du jury
Mme.	Fabienne	COMTE	Rapporteur
Mme.	Sylvie	HUET	Examineur
M.	Jean-Michel	LOUBES	Rapporteur
Mme.	Patricia	REYNAUD-BOURET	Examineur

Laboratoire Jean-Alexandre Dieudonné
Université de Nice, Parc Valrose
06108 NICE Cedex 2



Remerciements

Tout naturellement, mes premiers remerciements vont à Yannick pour avoir encadré cette thèse ainsi que mon travail pendant les années qui la précédèrent. La disponibilité dont tu as fait preuve et nos discussions, riches en conseils et en aide de ta part, m'ont beaucoup apporté en statistiques et m'ont appris à écrire. Au travers de tes qualités, tant humaines que scientifiques, j'ai découvert le monde de la recherche. Je profite donc de ces quelques lignes pour te dire ma reconnaissance, être ton élève fut, pour moi, un réel plaisir.

Je remercie Fabienne Comte et Jean-Michel Loubes qui ont directement accepté de rapporter ma thèse. Leurs commentaires furent instructifs et motivants et leurs présences dans le jury m'honnorent. Je suis particulièrement flatté que Lucien Birgé et Sylvie Huet fassent partie de mon jury ainsi que Patricia Reynaud-Bouret qui me donna mon premier véritable cours de statistiques et de nombreuses explications (même lorsque mes questions étaient posées depuis le sud de l'Inde).

La présence de Christophe Giraud et de Christine Mallot-Tuleau à Nice m'a également beaucoup apporté. Je les remercie pour leur sympathie et pour leur soutien.

Cette thèse doit aussi beaucoup à la qualité du cadre de travail offert par le laboratoire Dieudonné. J'ai une pensée particulière pour tous ceux sans qui la recherche serait un véritable cauchemar et qui, bien au-delà du "papier et des crayons", représentent une force indispensable. Merci donc à Rosalba, Isabelle, Stephanie, Janine, Jean-Marc, Isabelle, Julien, Jean-Paul, Fernande, Christiane, Jean-Louis, Claudine et Cecile.

Je sais la chance que j'ai eu, au cours de mes études, de recevoir d'excellents enseignements. Entre autres, je remercie Jean-François Le Gall qui m'a fait aimer les probabilités et Pascal Massart qui m'a fait découvrir avec enthousiasme la sélection de modèle. Je souhaite aussi remercier Sophie Schbath et l'unité MIG pour l'accueil chaleureux que j'ai reçu lors de ma visite à l'INRA de Jouy-en-Josas. Ce fut une chance de travailler dans de telles conditions et une expérience enrichissante tant au niveau scientifique que personnel.

Lors des conférences, séminaires et colloques, j'ai rencontré bien des personnes avec qui les échanges furent instructifs et amicaux. Pour tous ces bons moments, je remercie Adrien, Bertrand, Fanny, Katia, Matthieu, Mohamed, Sébastien et Sylvain. Bien sûr, je n'oublie pas Nicolas que j'ai la chance d'avoir comme ami depuis de nombreuses années.

La bonne ambiance qui a régné dans le bureau 617 durant ces dernières années a largement contribué au plaisir que j'ai eu à y travailler. Merci à celles et ceux qui ont animé ce lieu et qui l'animent encore: Ducduy, Nicolas, Hugues (qui a toujours des réponses), Asma, Luca et Chiara. Plus généralement, merci aux doctorants (et anciens doctorants) avec qui manger ou prendre un café a toujours été un plaisir: Benedikt, Daniel, Fabien, Hugo, Joan, José, Julianna, Laura, Marie, Michel, Nicolas, Nicolas, Patrick, Rémy, Salissou, Thomas et Thomas. De plus, je remercie tout spécialement ceux qui sont devenus des amis proches avec qui j'espère pouvoir continuer à partager tant de choses: Delphine, Fanny, Jean-Pascal, Marc, Marcello, Olivier, Philippe et Pierre.

Merci aux expatriés toulousains, Brigitte et Damien (et ses légions de dragons noirs), pour leur amitié et leur soutien. Merci aussi à Nico et Noé pour tout ce qu'ils représentent pour moi. Que mes

autres amis occitans se rassurent, ce n'est pas parce que je ne fais pas de liste exhaustive qu'ils sont oubliés. Ce temps passé à grandir ensemble reste mon meilleur réservoir à souvenirs.

Enfin, j'adresse de tendres remerciements à mes parents et à ma sœur pour tout ce qu'ils savent être important. Et Maud, pour cette indicible joie d'être deux.

Notations

If C is a constant, the notation $C(\cdot)$ specifies the dependency of C on some quantities.

w.r.t	with respect to
\square	end of a proof
\mathbb{N}	set of all nonnegative integers
\mathbb{R}	set of all real numbers
\mathbb{R}_+	set of all nonnegative real numbers
\mathbb{R}_+^*	set of all positive real numbers
\mathbb{M}_n	set of all real $n \times n$ -matrix
\mathbb{P}	probability measure
\mathbb{E}	expectation w.r.t. \mathbb{P}
Var	variance w.r.t. \mathbb{P}
$\mathbb{1}_A$	indicator function of the set A
$\text{Card}(A)$	cardinal of the set A
A^c	complementary of the set A
$\dim(E)$	dimension of the linear space E
E^\perp	orthogonal space to the linear space E
$\text{Span}\{v_1, \dots, v_k\}$	linear span of the vectors $v_1, \dots, v_k \in \mathbb{R}^n$
$\lfloor x \rfloor$	largest integer smaller or equal to x
$\log x$	natural logarithm of $x \in \mathbb{R}_+^*$
$ x $	absolute value of x
$x \vee y$	maximum of x and y
$x \wedge y$	minimum of x and y
x_+	positive part of x , <i>i.e.</i> $0 \vee x$
x_-	negative part of x , <i>i.e.</i> $0 \vee -x$
$\underset{x \in A}{\operatorname{argmin}} f(x)$	argument of the minimum of f on the set A
$\underset{x \in A}{\operatorname{argmax}} f(x)$	argument of the maximum of f on the set A
I_n	unit matrix of size n
$\mathbf{0}_{n,m}$	null matrix of size $n \times m$
$\text{Tr}(A)$	trace of the matrix A
$\text{rk}(A)$	rank of the matrix A
$\rho(A)$	spectral norm of the matrix A (see (3.1.16) in chapter 3)

tA	transpose of the matrix A
$\text{Im}(A)$	image set of the matrix A
$\ker(A)$	kernel set of the matrix A
$L^2(A, dx)$	set of functions $f : A \subset \mathbb{R} \rightarrow \mathbb{R}$ such that $\int_A f(x)^2 dx < \infty$
$(t_1, \dots, t_n)'$	transposed version of the vector $t \in \mathbb{R}^n$
$\ t\ $	L^2 -norm of $t \in \mathbb{R}^n$, <i>i.e.</i> $\ t\ = \left(\sum_{i=1}^n t_i^2 \right)^{1/2}$
$\ t\ _n$	normalized L^2 -norm of $t \in \mathbb{R}^n$, <i>i.e.</i> $\ t\ _n = \left(\frac{1}{n} \sum_{i=1}^n t_i^2 \right)^{1/2}$
$\langle t, t' \rangle_n$	normalized canonic scalar product of \mathbb{R}^n , <i>i.e.</i> $\langle t, t' \rangle_n = \frac{1}{n} \sum_{i=1}^n t_i t'_i$

Table des matières

Remerciements	i
Notations	iii
Chapitre 1. Introduction	1
1.1. Cadre de la régression	2
1.1.1. Régression hétéroscédastique	2
1.1.2. Modèle additif	3
1.2. Sélection de modèle	4
1.2.1. Motivations	5
1.2.2. Estimation par critère pénalisé	6
1.2.3. Propriétés de l'estimateur	7
1.3. Contributions de la thèse	10
Chapitre 2. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression	11
2.1. Introduction	11
2.2. Main results	14
2.2.1. Model collection and estimators	14
2.2.2. Risk upper bound	15
2.2.3. Convergence rate	16
2.3. Simulation study	17
2.4. Proofs	20
2.4.1. Proof of Proposition 2.1	20
2.4.2. Proof of Theorem 2.2	22
2.4.3. Proof of Proposition 2.3	29
2.5. Technical results	30
Chapitre 3. Estimation of a component in an additive model	33
3.1. Introduction	33
3.1.1. Additive models	33
3.1.2. Statistical framework	34
3.2. Main results	38
3.3. Application to additive models	40
3.4. Convergence rates	42
3.5. Estimation when σ^2 is unknown	44
3.6. Simulation study	45
3.6.1. Collections of models	45
3.6.2. The choice of C	47
3.6.3. Numerical simulations	48
3.6.4. Estimation of L^2 ratio	50
3.7. Proofs	53
3.7.1. Proofs of Theorems 3.2 and 3.3	53

3.7.2.	Proofs of Theorems 3.7 and 3.8	60
3.7.3.	Proofs of Corollaries and Propositions	63
3.8.	Lemmas	64
Appendice A.	Quadratic risk of the LSE and the MLE in a Gaussian framework with dependent data : an example	71
A.1.	Notations and recalls	71
A.2.	Study of the risk	72
Bibliographie		75

Introduction

“The time has come,” the Walrus said,
“to talk of many things.”

LEWIS CARROLL

Les dernières décennies ont connu un fort essor des interactions entre les statistiques mathématiques et de nombreux domaines de recherche (génétique, écologie, imagerie médicale, finance, ...). De la modélisation à l’analyse de données complexes, ces relations ont permis de meilleures approches de certains phénomènes. Cependant, elles ont aussi grandement étendu le champ de recherche des statisticiens. Ce lien entre les statistiques et le milieu des sciences appliquées trouve souvent sa source dans le souhait du non-statisticien d’avoir accès à des outils et des procédures lui permettant de formaliser son étude dans un cadre mathématique et de pouvoir traiter convenablement les données observées. Cela se traduit pour lui par la construction de modèles simples à expliquer mais assez riches pour prendre en compte la complexité du phénomène observé. C’est dans cette recherche de compromis entre adéquation aux données et relative simplicité du modèle que la théorie de la sélection de modèle trouve, entre autres, un vaste champ d’application.

Lors des interactions entre statisticiens et expérimentateurs, une des premières étapes consiste à prendre les probabilités comme langage commun. En particulier, ils doivent s’accorder sur une modélisation probabiliste satisfaisante pour les uns comme pour les autres. A la base de tout ce qui se construira ensuite, cette étape est cruciale et doit rendre compte de la connaissance du phénomène dont disposent les expérimentateurs.

Bien que les récents progrès des statistiques aient permis d’envisager des modèles de plus en plus généraux, le paramètre de variance des variables aléatoires introduites lors de cette étape de modélisation reste encore souvent supposé connu et constant. Dans le monde de l’expérimentateur, cette quantité est, malheureusement, bien souvent inaccessible et doit être approchée pour mettre en place les procédures proposées par le statisticien. Cela donne lieu à des méthodes utilisées en pratique mais dont les propriétés mathématiques peuvent être difficiles (voire impossibles) à établir. Ce constat fait apparaître la nécessité de prendre en compte la nature inconnue et potentiellement variable de la variance dans la construction de nouveaux outils statistiques.

Cette thèse se situe dans le cadre de la théorie statistique de la sélection de modèle. Elle propose une étude non-asymptotique de plusieurs problèmes liés à l’hétéroscédasticité. Les applications des résultats, obtenus dans des cadres généraux, sont axées sur l’estimation de paramètres en régression. Afin d’illustrer ces applications, des études de simulations sont réalisées à la fin de chaque chapitre. Cette introduction présente les principales idées développées dans la thèse. Le chapitre 2 traite de l’estimation simultanée de la moyenne et de la variance d’un vecteur gaussien à composantes indépendantes. Dans la suite de la thèse, nous nous intéressons au cas de données inter-dépendantes. Nous représentons la structure de dépendance des observations au moyen d’une matrice connue à un facteur multiplicatif σ près. Dans ce cadre, au chapitre 3, nous présentons des résultats sur l’estimation d’une composante dans un modèle additif pour σ connu ou inconnu.

1.1. Cadre de la régression

1.1.1. Régression hétéroscédastique. L'analyse des modèles régressifs est un sujet mathématique ancien. Les premiers travaux dans le domaine sont dus à Legendre [Leg05] et à Gauss [Gau09] pour l'estimation des orbites de certains corps du système solaire. Etant données deux variables $X \in \mathcal{X} \subset \mathbb{R}^d$ et $Y \in \mathcal{Y} \subset \mathbb{R}$, les modèles régressifs permettent d'expliquer les variations de Y en fonction de celles de X . De façon générale, ces modèles se présentent sous la forme

$$Y = s(X) + \sigma(X)\varepsilon \quad (1.1.1)$$

où ε est un terme de bruit (ou d'erreur), $s : \mathcal{X} \rightarrow \mathcal{Y}$ est appelée *fonction de régression* et $\sigma : \mathcal{X} \rightarrow \mathbb{R}_+^*$ s'appelle le *niveau de bruit*. La variable ε est supposée centrée et de variance unitaire conditionnellement à X (mais pas forcément indépendante de X), ce qui donne une autre définition de la fonction de régression,

$$s(x) = \mathbb{E}[Y|X = x], \quad x \in \mathcal{X}.$$

Ainsi, expliquer comment Y fluctue en fonction de X revient à déterminer des fonctions s et σ de telle sorte que (1.1.1) décrive au mieux la réalité du phénomène. D'un point de vue statistique, étant données des observations $X_1, \dots, X_n \in \mathcal{X}$ et $Y_1, \dots, Y_n \in \mathcal{Y}$, le problème consiste à construire de telles fonctions uniquement à partir des couples (X_i, Y_i) . Nous supposons donc qu'il existe deux fonctions s et σ inconnues telles que

$$Y_i = s(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \dots, n, \quad (1.1.2)$$

et l'objectif est d'en donner des *estimateurs* construits à partir des observations. Lorsque les X_i sont des variables aléatoires, ce problème de régression est dit à *support aléatoire*. Inversement, lorsque les X_i sont déterministes et connues, nous parlons de *support fixe*. Les résultats présentés dans la suite de cette thèse sont tous établis dans le cadre de la régression à support fixe. Pour plus de précisions sur ces deux situations, le lecteur pourra consulter [Bar00] et [Bar02].

Désormais, les variables du support seront notées en lettres minuscules afin de garder à l'esprit leur nature déterministe. Considérons donc les observations $(x_1, Y_1), \dots, (x_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ et supposons, dans un premier temps, qu'elles sont indépendantes et que le niveau de bruit $\sigma(\cdot) \equiv \sigma > 0$ est constant. Un tel cas est appelé *régression homoscedastique*. Il est possible de reformuler (1.1.2),

$$Y_i = s_i + \sigma\varepsilon_i, \quad i = 1, \dots, n, \quad (1.1.3)$$

où $s_i = s(x_i)$ et $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ est un vecteur aléatoire dont les composantes sont indépendantes, centrées et de variance 1. L'estimation de la fonction s à partir de (1.1.3), que σ soit connue ou

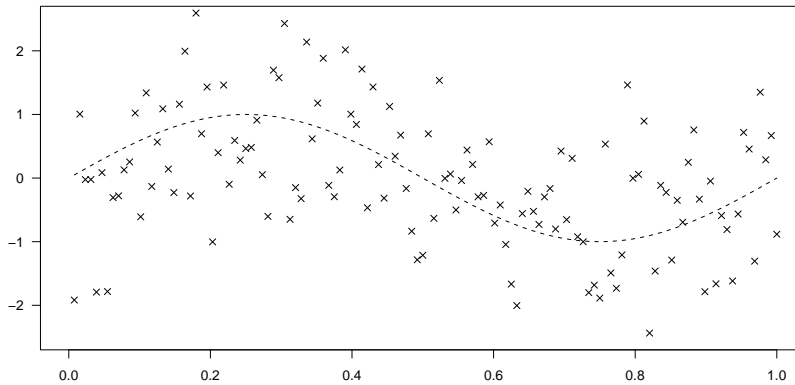


FIGURE 1.1. Exemple de données d'une régression homoscedastique gaussienne à coordonnées indépendantes ($n = 128$, $\mathcal{X} = [0, 1]$, $s(x) = \sin(2\pi x)$, $\sigma^2(x) = 1$)

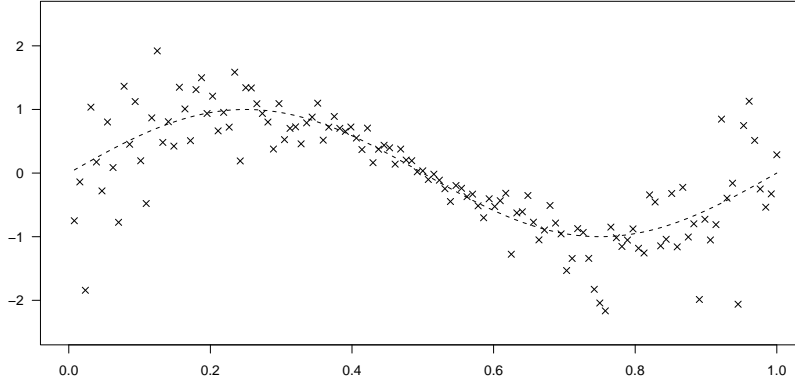


FIGURE 1.2. Exemple de données d'une régression hétéroscédastique gaussienne à coordonnées indépendantes ($n = 128$, $\mathcal{X} = [0, 1]$, $s(x) = \sin(2\pi x)$, $\sigma^2(x) = 1 - 3.96x(1 - x)$)

inconnue, a été le sujet de nombreux travaux. Dans un cadre asymptotique (*i.e.* pour un nombre d'observations n tendant vers l'infini), citons, par exemple, les travaux de Shibata [Shi81] dans le cas où les ε_i sont gaussiens, Li [Li87] si les ε_i admettent un moment d'ordre 8 et Polyak et Tsybakov [PT90] pour un moment d'ordre 4 seulement. Le cadre non-asymptotique (*i.e.* pour n fixe donné) a été étudié, entre autres, par Barron, Birgé et Massart [BBM99] pour le cas gaussien et par Baraud [Bar00] sous l'hypothèse que les ε_i admettent un moment d'ordre plus grand que 2.

Nous parlons de *régression hétéroscédastique* dès lors que le niveau de bruit σ n'est plus supposé constant. Dans ce cas, (1.1.2) devient

$$Y_i = s_i + \sigma_i \varepsilon_i, \quad i = 1, \dots, n, \quad (1.1.4)$$

où les $\sigma_i = \sigma(x_i)$ dépendent maintenant du support des observations et sont, a priori, inconnus. Les résultats théoriques d'estimation des fonctions s et/ou σ dans le modèle (1.1.4) sont peu nombreux et relativement récents. Dans un cadre gaussien, Comte et Rozenholc [CR02] estiment s puis σ dans une procédure de sélection de modèle en deux étapes et Galtchouk et Pergamenschikov [GP05] ont construit un estimateur adaptatif de s avec σ inconnue. Dans le chapitre 6 de [Arl07], Arlot propose une procédure d'estimation de s dans un modèle hétéroscédastique pour des données bornées. Enfin, de récents travaux de Cai et Wang [CW08] et Wang [Wt08] ont porté sur l'estimation de σ avec s inconnue ainsi que des effets de la régularité de s sur cette estimation. Une discussion plus étendue sur le modèle (1.1.4) ainsi qu'une procédure d'estimation de s et σ feront l'objet du chapitre 2 de cette thèse.

1.1.2. Modèle additif. Afin de décrire la façon dont Y varie en fonction de la *variable explicative* $x = (x^{(1)}, \dots, x^{(d)})' \in \mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_d \subset \mathbb{R}^d$, un modèle particulier est depuis longtemps étudié, celui de la *régression linéaire*. Dans ce cadre, les observations se mettent sous la forme

$$Y_i = \beta_0 + \sum_{j=1}^d \beta_j x_i^{(j)} + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

où les $\beta_j \in \mathbb{R}$ sont inconnus. Ce modèle présente l'avantage d'être simple à interpréter en pratique et permet une analyse composante par composante des effets de x sur Y . Cependant, ce modèle demeure trop simple dans de nombreux cas et ne permet pas de modéliser des relations plus complexes que la seule linéarité. Afin de palier à ce manque de flexibilité, on considère des *modèles additifs* pour

lesquels les observations sont

$$Y_i = \beta_0 + \sum_{j=1}^d f_j(x_i^{(j)}) + \sigma \varepsilon_i, \quad i = 1, \dots, n, \quad (1.1.5)$$

où les fonctions inconnues $f_j : \mathcal{X}_j \rightarrow \mathbb{R}$ sont appelées les *composantes* de la fonction de régression. Les premiers travaux dans ce cadre furent ceux de Leontief [Leo47] et de Scheffé [Sch59] qui qualifiaient ces modèles de *séparables additifs* et *additifs sans interaction* respectivement. Ces modèles sont, aujourd'hui encore, très répandus en économie théorique et en économétrie car ils y jouissent d'une grande interprétabilité. Pour de nombreux exemples d'applications des modèles additifs à l'économie, on pourra consulter les références données en fin du chapitre 8 de [HMSW04].

Pour le statisticien, étudier ces modèles signifie estimer les composantes f_j à partir des observations $(x_1, Y_1), \dots, (x_n, Y_n)$. Depuis les travaux de Stone [Sto85], il est bien connu que la vitesse de convergence optimale pour estimer la fonction de régression s dans un cadre général homoscédastique (1.1.3) est de l'ordre de $n^{-\alpha/(2\alpha+d)}$ où $\alpha > 0$ est un indice de régularité de s . Cette vitesse est d'autant plus faible que la dimension d de la variable explicative est grande (ce phénomène est appelé *fléau de la dimension*). Stone [Sto85] a montré que dans un cadre additif (1.1.5), la vitesse optimale d'estimation d'une composante f_j est de l'ordre de la vitesse unidimensionnelle $n^{-\alpha/(2\alpha+1)}$. En d'autres termes, l'estimation de la composante f_j à partir de (1.1.5) peut être faite avec la même vitesse optimale que celle atteignable par une procédure construite à partir des observations $Y'_i = f_j(x_i^{(j)}) + \sigma \varepsilon_i$. Ce fait notable a grandement motivé les études statistiques d'estimation de composantes dans un modèle additif depuis le début des années 80. Les travaux de Buja *et al.* [BHT89], Hastie et Tibshirani [HT90], en particulier, ont largement contribué à ce domaine. Il existe deux approches populaires pour estimer ces composantes, le *backfitting* et l'*intégration marginale*. Les résultats théoriques obtenus par l'une ou par l'autre sont de nature asymptotique et il existe peu de résultats non-asymptotiques sur le sujet. En utilisant des méthodes de sélection de modèle, Baraud [Bar02], Comte et Viennet [BCV01] ont proposé des méthodes d'estimation de la fonction de régression dans des modèles additifs sous des hypothèses de moment sur les ε_i . Plus récemment, Brunel et Comte (voir [BC06] et [BC08]) ont obtenu des résultats similaires pour des modèles additifs censurés.

Dans le chapitre 3 de cette thèse, nous présentons une nouvelle méthode d'estimation non-asymptotique d'une composante dans un modèle additif. Le principe de celle-ci consiste à projeter convenablement les observations (1.1.5) sur un sous-espace de \mathbb{R}^n de façon à réduire l'impact des composantes supplémentaires. Idéalement, pour estimer la composante f_1 par exemple, l'opération correspondrait à la donnée d'une matrice de projection P telle que $\mathbb{E}[(PY)_i] = f_1(x_i^{(1)})$. Les composantes étant inconnues, une telle matrice n'est pas disponible pour le statisticien. En revanche, il est possible de construire explicitement une matrice ayant des propriétés similaires à partir des points du support. Plus généralement, si P est une matrice quelconque, nous établissons des résultats de sélection de modèle dans le cadre régressif donné par

$$(PY)_i = Y'_i = s'_i + \sigma (P\varepsilon)_i, \quad i = 1, \dots, n. \quad (1.1.6)$$

Notons, en particulier, que ces observations hétéroscédastiques ne sont pas indépendantes contrairement au cas (1.1.4).

Les schémas de régression (1.1.3), (1.1.4) et (1.1.6) admettent, par définition, des structures probabilistes différentes. Les figures 1.1, 1.2 et 1.3 illustrent cette différence de nature entre les phénomènes mis en jeu. La prise en compte de ces structures pour la construction de procédures d'estimation fut un des principaux objectifs de cette thèse.

1.2. Sélection de modèle

Les travaux présentés dans cette thèse sont basés sur des méthodes de sélection de modèle dans un cadre non-asymptotique. Afin d'introduire les concepts de base de cette théorie, nous présentons ici la problématique de la sélection de modèle dans le cadre de la régression à composantes indépendantes.

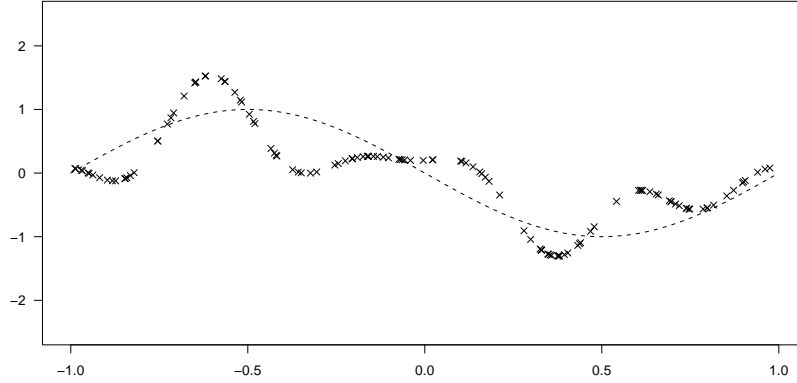


FIGURE 1.3. Exemple de données d'une régression hétéroscédastique gaussienne à coordonnées dépendantes : projection des observations sur un espace trigonométrique ($n = 128$, $\mathcal{X} = [-1, 1]$, $s(x) = -\sin(\pi x)$, $\sigma^2 = 1$)

Nos principales références pour cette introduction sont le cours de Saint-Flour de Massart [Mas07] et l'introduction de la thèse d'Arlot [Arl07].

Par souci de lisibilité, nous identifions implicitement les fonctions s et σ aux vecteurs $(s_1, \dots, s_n)'$ et $(\sigma_1, \dots, \sigma_n)'$ respectivement. De plus, nous notons génériquement $\theta \in \Theta$ le paramètre à estimer qui pourra être s ($\Theta = \mathbb{R}^n$) ou (s, σ) ($\Theta = \mathbb{R}^n \times \mathbb{R}^n$) selon le cas dans lequel nous nous placerons.

1.2.1. Motivations. Un *contraste empirique* γ_n pour l'estimation de θ est une fonction réelle définie sur Θ à partir des observations Y_i telle que

$$\theta' \in \Theta \mapsto \mathbb{E}[\gamma_n(\theta')]$$

admette un minimum en θ . Etant donné un espace linéaire $S \subset \Theta$ que nous appelons *modèle*, une approche classique pour estimer θ dans S est de définir l'estimateur $\hat{\theta}$ de θ comme un minimiseur de γ_n sur S . A tout contraste empirique, il est possible d'associer une *fonction de perte* définie par

$$\ell(\theta, \theta') = \mathbb{E}[\gamma_n(\theta')] - \mathbb{E}[\gamma_n(\theta)] \geq 0, \quad \theta' \in S.$$

Pour l'estimation de $\theta = s$ dans le cadre homoscedastique (1.1.3), un contraste populaire est celui des moindres carrés,

$$\gamma_n(t) = \sum_{i=1}^n (Y_i - t_i)^2, \quad t \in \mathbb{R}^n,$$

et sa fonction de perte associée vaut

$$\ell(s, t) = \|s - t\|^2 = \sum_{i=1}^n (s_i - t_i)^2. \quad (1.2.1)$$

Dans ce cas, l'estimateur de s est appelé *estimateur des moindres carrés* et correspond à la projection orthogonale de Y sur S ,

$$\hat{s} = \pi Y = \operatorname{argmin}_{t \in S} \gamma_n(t).$$

Si les termes de bruit ε_i sont gaussiens, la *fonction de vraisemblance* peut être utilisée comme contraste pour estimer $\theta = (s, \sigma)$ dans une régression hétéroscédastique (1.1.4),

$$\gamma_n(t, \tau) = \frac{1}{2} \sum_{i=1}^n \frac{(Y_i - t_i)^2}{\tau_i^2} + \ln \sqrt{2\pi\tau_i^2}, \quad t \in \mathbb{R}^n, \quad \tau \in (0, \infty)^n.$$

La fonction de perte associée est la *divergence de Kullback-Leibler*

$$\ell((s, \sigma), (t, \tau)) = \frac{1}{2} \sum_{i=1}^n \frac{(s_i - t_i)^2}{\tau_i^2} + \ln \frac{\tau_i^2}{\sigma_i^2} + \frac{\sigma_i^2}{\tau_i^2} - 1 \quad (1.2.2)$$

et l'estimateur $\hat{\theta} = (\hat{s}, \hat{\sigma}^2) \in S$ obtenu par minimisation de γ_n sur S est l'*estimateur du maximum de vraisemblance*.

L'heuristique motivant le choix d'un estimateur par minimisation de contraste empirique consiste à dire qu'un minimiseur de $\gamma_n(\theta')$ sur S ne doit pas être trop “éloigné” de celui de $\mathbb{E}[\gamma_n(\theta')]$, au moins dans le cas où θ appartient à S . En d'autres termes, nous espérons que le *risque* $\mathbb{E}[\ell(\theta, \hat{\theta})]$ de $\hat{\theta}$ soit relativement petit. Dans les situations qui nous intéressent ici (et sous certaines hypothèses dans le cadre hétéroscédastique), ce risque est de l'ordre de la somme

$$\inf_{\theta' \in S} \ell(\theta, \theta') + \sigma^2 D$$

où D est la dimension de S . Le premier terme, dit de *biais*, représente la capacité de S à approcher θ . Le second, dit de *variance*, est proportionnel à D et correspond à la “taille” du modèle. La quantité θ étant inconnu, nous pourrions être tenter de prendre un “gros” modèle S (*i.e.* de grande dimension) afin d'espérer avoir un petit terme de biais mais cela donnerait un bien mauvais risque du fait de la taille du terme de variance. A l'opposé, choisir un petit modèle S (une constante par exemple, $D = 1$) assure une faible variance mais un mauvais biais dès lors que θ est trop éloigné de S . Nous comprenons ainsi pourquoi choisir un “bon” modèle revient à trouver un compromis entre le terme de biais et celui de variance.

Afin d'illustrer ce problème à la base de la sélection de modèle, considérons l'estimation de s dans le cadre de la régression homoscédastique (1.1.3) pour un modèle S_m d'histogrammes sur une partition régulière m de $\{1, \dots, n\}$ à D_m blocs de même taille,

$$S_m = \left\{ \sum_{\lambda \in m} \alpha_\lambda \mathbb{1}_\lambda : \forall \lambda \in m, \alpha_\lambda \in \mathbb{R} \right\}. \quad (1.2.3)$$

L'estimateur des moindres carrés de s s'écrit alors

$$\hat{s}_m = \sum_{\lambda \in m} \hat{\alpha}_\lambda \mathbb{1}_\lambda \quad \text{avec} \quad \hat{\alpha}_\lambda = \frac{1}{\text{Card}(\lambda)} \sum_{i \in \lambda} Y_i$$

et son risque vaut

$$\mathbb{E} [\|s - \hat{s}_m\|^2] = \inf_{t \in S_m} \|s - t\|^2 + \sigma^2 D_m. \quad (1.2.4)$$

La figure 1.4 montre trois choix possibles pour la partition m . Si $D_m = 1$, le modèle n'est visiblement pas assez riche pour approcher s et, pour $D_m = 42$, l'estimateur donne trop d'importance aux données et explique s par du bruit. Visuellement, le cas $D_m = 8$ semble être un bon candidat pour le compromis biais-variance abordé précédemment.

1.2.2. Estimation par critère pénalisé. Considérons, plus généralement, une collection de modèles $\{S_m, m \in \mathcal{M}\}$ au plus dénombrable et un contraste empirique γ_n . Pour chaque $m \in \mathcal{M}$, nous définissons l'estimateur $\hat{\theta}_m$ comme un minimiseur de γ_n sur S_m , D_m la dimension de S_m et π_m la projection orthogonale sur S_m . Suivant la problématique décrite dans la sous-section précédente, nous souhaitons construire une procédure basée sur les observations Y_i qui nous permette de choisir un estimateur parmi la collection $\{\hat{\theta}_m, m \in \mathcal{M}\}$ qui ait un risque minimal.

Parmi \mathcal{M} , il existe un indice $m(\theta) \in \mathcal{M}$ tel que

$$m(\theta) \in \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}[\ell(\theta, \hat{\theta}_m)].$$

La variable $\hat{\theta}_{m(\theta)}$ est appelé l'*oracle* par référence à [DJ94]. Notons que $\hat{\theta}_{m(\theta)}$ n'est pas un estimateur de θ car le choix de $m(\theta)$ dépend de θ lui-même. Cependant, son risque étant minimal parmi ceux

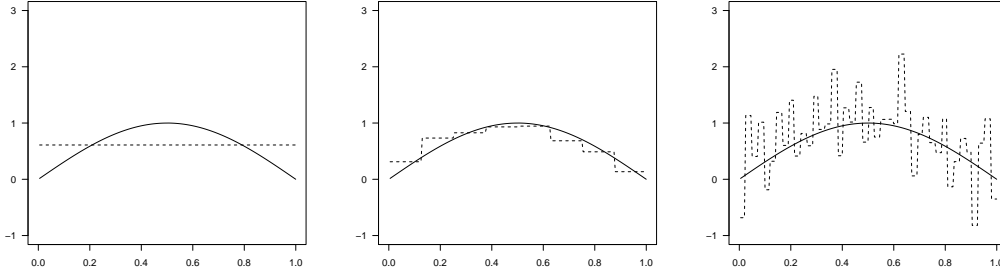


FIGURE 1.4. Estimation de s par un histogramme régulier sur 1, 8 et 42 blocs de gauche à droite ($n = 200$, $s(x) = \sin(\pi x)$ et $\sigma^2 = 1$)

des estimateurs $\hat{\theta}_m$, il servira de référence à notre procédure de sélection de modèle, l'objectif étant de choisir un estimateur dont le risque est aussi proche que possible de celui de l'oracle.

Pour faire ce choix, nous procédons par *pénalisation*. Soit une *fonction de pénalité* $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+$, nous définissons l'indice $\hat{m} \in \mathcal{M}$ comme un minimiseur du *critère pénalisé*

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \gamma_n(\hat{\theta}_m) + \text{pen}(m) \right\}, \quad (1.2.5)$$

et nous posons $\tilde{\theta} = \hat{\theta}_{\hat{m}}$. Les premiers résultats obtenus par critère pénalisé sont dus à Akaike [Aka73] pour l'estimation de densité par vraisemblance pénalisée et à Mallows [Mal73] pour l'estimation de la fonction de régression dans un cadre gaussien homoscedastique (1.1.3) à variance σ^2 connue. Dans les deux cas, l'heuristique consiste à prendre une pénalité de l'ordre du risque de $\hat{\theta}_m$. Pour $\text{pen}_{\text{Akaike}}(m) = D_m$, Akaike obtint le critère AIC et, pour $\text{pen}_{\text{Mallows}}(m) = 2\sigma^2 D_m$, le critère s'appelle C_p de Mallows. Shibata [Shi81] et Birgé et Massart [BM01a] ont montré que ces critères sont asymptotiquement optimaux à condition que la taille de \mathcal{M} ne soit pas trop grande. Dans un cadre non-asymptotique, Birgé et Massart [BM01a] ont introduit des pénalités plus générales que celle de Mallows et ont obtenu un contrôle sur le risque de $\tilde{\theta}$ quelque soit la taille de la collection de modèle. La construction de leur procédure se base sur des résultats de concentration de la mesure gaussienne.

À l'exception de AIC, ces critères nécessitent la connaissance de la variance σ^2 dans (1.1.3) et ne sont plus directement utilisables dans le cas où celle-ci est inconnue. En pratique, une solution consiste à remplacer σ^2 dans la pénalité par un estimateur $\hat{\sigma}^2$. Apparaît alors le problème d'estimation de σ^2 : ne connaissant pas, a priori, de "bon" modèle pour estimer s , lequel choisir pour estimer la variance ? Récemment, Baraud, Giraud et Huet [BGH09] ont étudié la minimisation du critère suivant,

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - \pi_m Y\|^2 (1 + \text{pen}(m)) \right\}, \quad (1.2.6)$$

afin d'estimer s à partir de (1.1.3) avec un bruit gaussien de variance inconnue. Notons que, dans ce critère, la pénalité est multiplicative et $\|Y - \pi_m Y\|^2$ sert d'estimateur de la variance dans le modèle S_m . Baraud *et al.* ramènent des critères classiques tels que FPE (voir [Aka70]), AIC ou BIC (voir [Sch78]) à une forme (1.2.6) et, dans chacun des cas, étudient d'un point de vue non-asymptotique le risque quadratique de $\pi_{\hat{m}} Y$ en fonction de la taille de \mathcal{M} . En se basant sur des quantiles de Fisher, ils introduisent aussi de nouvelles pénalités capables de prendre en compte la taille de la collection de modèles. Pour l'estimation non-asymptotique de s en régression hétéroscédastique (1.1.4), Arlot a donné des pénalités construites par rééchantillonnage mais leurs descriptions sortent du cadre de cette introduction.

1.2.3. Propriétés de l'estimateur. La qualité d'une procédure de sélection de modèle correspond à sa capacité à choisir un estimateur $\tilde{\theta}$ parmi une collection $\{\hat{\theta}_m, m \in \mathcal{M}\}$ qui ait un risque faible. Afin d'évaluer cette qualité, nous pouvons considérer deux propriétés : les inégalités de type

oracle et l'adaptativité. Nous présentons maintenant ces notions dans le cas où la fonction de perte est celle des moindres carrés (1.2.1) ou la divergence de Kullback-Leibler (1.2.2).

Afin de valider théoriquement la qualité de l'estimateur $\tilde{\theta}$, nous chercherons à établir des inégalités de la forme

$$\mathbb{E} [\ell(\theta, \tilde{\theta})] \leq C \inf_{m \in \mathcal{M}} \left\{ \inf_{\theta' \in S_m} \ell(\theta, \theta') + \text{pen}(m) \right\} + R$$

où C est une constante universelle et R un terme de reste. Pour une pénalité linéaire en la dimension du modèle, la quantité dans l'infimum est de l'ordre du risque de l'estimateur $\hat{\theta}_m$. Ainsi, de telles inégalités permettent de comparer le risque de $\tilde{\theta}$ à celui de l'oracle $\hat{\theta}_{m(\theta)}$ qui sert de référence à notre procédure. Sous certaines hypothèses sur la collection de modèles, ce type de résultat donne lieu à des inégalités dites *oracles*,

$$\mathbb{E} [\ell(\theta, \tilde{\theta})] \leq C \inf_{m \in \mathcal{M}} \mathbb{E} [\ell(\theta, \hat{\theta}_m)] = C \mathbb{E} [\ell(\theta, \hat{\theta}_{m(\theta)})] .$$

Reprenons l'exemple de l'estimation de s en régression gaussienne (1.1.3) à variance connue et constante. Birgé et Massart [BM01a] ont montré que si la pénalité est de la forme

$$\text{pen}(m) = K\sigma^2 D_m \left(1 + \sqrt{2L_m}\right)^2, \quad m \in \mathcal{M},$$

avec $K > 1$ et $\{L_m, m \in \mathcal{M}\}$ une collection de poids positifs, alors l'estimateur \tilde{s} obtenu par minimisation du critère des moindres carrés pénalisé est tel que

$$\mathbb{E} [\|s - \tilde{s}\|^2] \leq C \inf_{m \in \mathcal{M}} \{ \|s - \pi_m s\|^2 + \text{pen}(m) \} + \sigma^2 \Sigma \quad (1.2.7)$$

où C ne dépend que de K et le terme de reste est donné par

$$\Sigma = \sum_{m \in \mathcal{M}} \exp(-L_m D_m) .$$

Cette quantité dépend de la collection de modèles et des poids L_m . Pour déduire une inégalité oracle à partir de (1.2.7), il faut donc exhiber une majoration générale de Σ . Pour $D \in \mathbb{N}$, nous notons N_D le nombre de modèles de dimension D dans la collection,

$$N_D = \text{Card} \{m \in \mathcal{M} : D_m = D\} .$$

Le seul modèle de dimension nulle étant $\{0\}$, nous supposons qu'il n'est pas présent plusieurs fois dans la collection de modèles et donc que $N_0 \leq 1$. En prenant des poids $\{L_D, D \in \mathbb{N}\}$ identiques pour les modèles de même dimension, nous obtenons

$$\Sigma \leq \sum_{D \in \mathbb{N}} N_D \exp(-L_D D) = 1 + \sum_{D > 0} \exp\left(-\left[L_D - \frac{\ln N_D}{D}\right] D\right) \mathbb{1}_{N_D > 0} .$$

Pour majorer la série indépendamment de la forme des modèles, il suffit de prendre des poids vérifiant, par exemple, $L_D > 2(\ln(N_D)/D) \mathbb{1}_{N_D > 0}$, pour tout $D > 0$. Notons que la pénalité dépend alors de la taille de la collection de modèles (*i.e.* du fait qu'elle contienne plus ou moins de modèles de même dimension). Si, pour tout $D > 0$, $N_D \leq 1$ (c'est le cas de la sélection de variables ordonnées, voir le chapitre 4 de [Mas07]), la collection est petite et nous pouvons prendre tous les L_D égaux à une constante $L > 0$. En faisant de sorte que $K(1 + \sqrt{2L})^2 = 2$, on retrouve d'ailleurs ainsi le critère de Mallows. Dans ce cas, Σ est majoré par $(e^L - 1)^{-1}$ et l'inégalité (1.2.7) mène à

$$\begin{aligned} \mathbb{E} [\|s - \tilde{s}\|^2] &\leq C \inf_{m \in \mathcal{M}} \{ \|s - \pi_m s\|^2 + 2\sigma^2 D_m \} + \frac{\sigma^2}{e^L - 1} \\ &\leq C' \inf_{m \in \mathcal{M}} \{ \|s - \pi_m s\|^2 + \sigma^2 (D_m \vee 1) \} . \end{aligned}$$

Pour les modèles de dimension non-nulle, la quantité dans l'infimum est le risque de \hat{s}_m et cette inégalité est donc de type oracle. Dans le cas d'une plus grande collection de modèles, il se peut que $\ln(N_D)/D$ ne soit pas majoré par un terme indépendant de la taille n de l'échantillon. Considérons le problème de la sélection de variables complète. Le nombre N_D de modèles de dimension D vaut

$n!/(D!(n-D)!) \simeq n^D$ et, pour majorer Σ , nous sommes donc amenés à prendre des poids L_D de l'ordre de $\ln n$. Le coefficient de D_m dans la pénalité étant, a priori, grand, les modèles de grandes dimensions sont défavorisés dans la procédure de choix de modèle (on parle de phénomène de *sur-pénalisation*). Avec ce choix des poids, l'inégalité (1.2.7) donne alors une inégalité de type oracle à un facteur logarithmique près

$$\mathbb{E} [\|s - \tilde{s}\|^2] \leq C' \inf_{m \in \mathcal{M}} \{ \|s - \pi_m s\|^2 + \sigma^2 (D_m \vee 1) \ln n \} .$$

Cependant, nous savons que ce facteur est inévitable dans le cas de la sélection de variables complète (voir le chapitre 4 de [Mas07] pour une plus longue discussion sur le sujet).

Les inégalités de type oracle présentent l'avantage d'être valables quelque soit la véritable valeur de θ . Néanmoins, elles ont aussi l'inconvénient de comparer le risque de $\tilde{\theta}$ uniquement avec les risques des estimateurs de la collection $\{\hat{\theta}_m, m \in \mathcal{M}\}$. Nous aimerions aussi pouvoir comparer ce risque avec d'autres estimateurs de θ . Naturellement, nous souhaitons des comparaisons relativement uniformes en θ puisque une constante θ_0 est toujours un estimateur (parfait si $\theta = \theta_0$ et très mauvais sinon). Une méthode classique pour cela est de considérer le risque maximal sur certains espaces \mathcal{T}_α caractérisés par une propriété dépendante d'un paramètre $\alpha \in A$ (la régularité de la fonction de régression, par exemple). Il s'agit du point de vue minimax: un estimateur est "bon" si son risque maximal sur \mathcal{T}_α est proche du *risque minimax* défini par

$$\mathcal{R}_\infty(\mathcal{T}_\alpha, \ell) = \inf_T \sup_{\theta \in \mathcal{T}_\alpha} \mathbb{E} [\ell(\theta, T)]$$

où l'infimum porte sur tous les estimateurs de θ (en particulier, ceux dépendants de α). La qualité de l'estimateur $\tilde{\theta}$ est donc mesurée par le rapport

$$\sup_{\theta \in \mathcal{T}_\alpha} \frac{\mathbb{E} [\ell(\theta, \tilde{\theta})]}{\mathcal{R}_\infty(\mathcal{T}_\alpha, \ell)}$$

et plus il est proche de un, meilleur est $\tilde{\theta}$. Nous dirons que l'estimateur $\tilde{\theta}$ est *adaptatif* (au sens du minimax) au paramètre α si, pour tout $\alpha \in A$, ce rapport est borné indépendamment de n . En d'autres termes, $\tilde{\theta}$ est adaptatif à α si

$$\forall \alpha \in A, \exists C_\alpha > 1 \text{ tel que } \sup_{\theta \in \mathcal{T}_\alpha} \mathbb{E} [\ell(\theta, \tilde{\theta})] \leq C_\alpha \mathcal{R}_\infty(\mathcal{T}_\alpha, \ell)$$

où C_α est un facteur numérique pouvant dépendre de certains paramètres tels que α ou σ^2 mais pas de n .

Dans le cadre de la régression, un avantage connu des inégalités oracles est de permettre relativement aisément d'obtenir des propriétés d'adaptativité au sens du minimax à la régularité de la fonction de régression dès lors que la collection de modèles a une bonne capacité d'approximation (voir [BM97]). Prenons le cas de la régression gaussienne homoscédastique sur le support fixe donné par les points $i/n, i = 1, \dots, n$. Le cadre statistique est (1.1.3) où les éléments s_i sont les valeurs de la fonction de régression, notée s par identification, aux points i/n . Pour $D_m \in \{1, \dots, n\}$, considérons le modèle S_m des histogrammes construits sur la partition régulière de $[0, 1]$ à D_m blocs de même taille. Puisque, pour tout $D \in \mathbb{N}$, $N_D \leq 1$, nous savons que par minimisation d'un critère de Mallows, nous pouvons obtenir une inégalité oracle pour le risque quadratique normalisé,

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] = \frac{\mathbb{E} [\|s - \tilde{s}\|^2]}{n} \leq C \inf_{1 \leq D_m \leq n} \left\{ \|s - \pi_m s\|_n^2 + \frac{\sigma^2 D_m}{n} \right\} \quad (1.2.8)$$

où C est une constante numérique ne dépendant pas de n . Pour $\alpha \in]0, 1[$, considérons la boule unité de Hölder de coefficient α ,

$$\mathcal{H}_\alpha = \{f : [0, 1] \rightarrow \mathbb{R} : \forall x, y \in [0, 1], |f(x) - f(y)| \leq |x - y|^\alpha\} .$$

La capacité d'approximation des fonctions de cet ensemble par les éléments de S_m est bien connue (voir [DL93]) et, en particulier, si $s \in \mathcal{H}_\alpha$, nous avons

$$\|s - \pi_m s\|_n^2 \leq D_m^{-2\alpha}.$$

Ainsi, pour tout $\alpha \in]0, 1[$, l'inégalité (1.2.8) nous donne

$$\begin{aligned} \sup_{s \in \mathcal{H}_\alpha} \mathbb{E} [\|s - \tilde{s}\|_n^2] &\leq C \inf_{1 \leq D_m \leq n} \left\{ D_m^{-2\alpha} + \frac{\sigma^2 D_m}{n} \right\} \\ &\leq C \left(\left[n^{1/(1+2\alpha)} \right]^{-2\alpha} + \frac{\sigma^2 \lfloor n^{1/(1+2\alpha)} \rfloor}{n} \right) \\ &\leq C_\alpha (1 + \sigma^2) n^{-2\alpha/(1+2\alpha)} \\ &= C_\alpha (1 + \sigma^2) \mathcal{R}_\infty(\mathcal{H}_\alpha, \|\cdot\|_n^2) \end{aligned}$$

où $\lfloor x \rfloor$ est la partie entière de $x \in \mathbb{R}$ et $C_\alpha > 1$ ne dépend que du paramètre α . L'estimateur \tilde{s} construit par cette procédure est donc adaptatif au coefficient de Hölder de la fonction de régression.

1.3. Contributions de la thèse

Le chapitre 2 est consacré à l'estimation simultanée de la moyenne s et de la variance σ dans le cadre de la régression hétéroscédastique (1.1.4) à partir de deux copies indépendantes d'un vecteur gaussien Y . Ce chapitre a fait l'objet d'une publication [Gen08] dans *Electronic Journal of Statistics*. Pour chaque m dans un ensemble d'indices \mathcal{M} , nous considérons un modèle $S_m \times \Sigma_m$ où S_m et Σ_m sont des espaces linéaires dévolus à l'estimation des vecteurs s et σ respectivement. Le couple (s, σ) est estimé par les estimateurs $(\hat{s}_m, \hat{\sigma}_m)$ du maximum de vraisemblance légèrement modifiés de telle façon que l'estimation de la moyenne soit indépendante de celle de la variance dans chaque modèle. Notons σ^* (resp. σ_*) le supremum (resp. infimum) de σ sur le support des observations. En supposant connue une quantité $\gamma > 1$ telle que $\sigma^*/\sigma_* \leq \gamma$, nous proposons une méthode de sélection de modèle basée sur un critère de vraisemblance pénalisé. Une majoration non-asymptotique du risque de Kullback de l'estimateur sélectionné est fournie ainsi que des vitesses de convergence sur les boules de Hölder dans le cas de la régression. Ces résultats ainsi que les performances de l'estimateur sont illustrés par plusieurs simulations à la section 2.3.

A partir du chapitre 3, nous nous concentrons sur la régression hétéroscédastique basée sur des observations inter-dépendantes. A partir d'un jeu de donnée (1.1.6), nous construisons une procédure de sélection de modèle non-asymptotique pour estimer une composante dans un modèle additif (1.1.5) à variance σ^2 connue ou non. Nous proposons des pénalités similaires à celles de Birgé et Massart [BM01a]. En minimisant un critère des moindres carrés pénalisé, nous obtenons des majorations du risque quadratique de notre estimateur sous l'hypothèse d'un bruit gaussien et sous une condition de moment semblable à celle considérée par Baraud [Bar00]. S'en déduisent des vitesses de convergence et des résultats d'adaptativité de nos estimateurs sur les boules de Hölder. Des simulations sont proposées afin de décrire les performances de nos estimateurs en pratique.

Le lecteur pourra trouver l'intégralité des codes sources des simulations ainsi qu'une bibliothèque C relative aux procédures statistiques du chapitre 3 sur la page de l'auteur. Celles du chapitre 2 sont basées sur le logiciel R¹, les autres utilisent la bibliothèque scientifique GSL².

¹voir [R D07]

²voir [GSL]

Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression

RÉSUMÉ. Considérons un vecteur gaussien Y dans \mathbb{R}^n de moyenne s et de matrice de covariance Γ diagonale. Dans ce chapitre, notre objectif est d'estimer s et les quantités $\sigma_i = \Gamma_{i,i}$, pour $i = 1, \dots, n$, à partir de l'observation de deux copies indépendantes de Y . Notre approche ne nécessite aucune hypothèse sur s mais demande la connaissance d'une borne supérieure γ sur le rapport $\max_i \sigma_i / \min_i \sigma_i$. Par exemple, le choix $\gamma = 1$ correspond au cas homoscédastique où les composantes de Y sont supposées avoir la même variance inconnue. D'un autre côté, le choix $\gamma > 1$ correspond au cas hétéroscédastique dans lequel les variances des composantes de Y peuvent varier dans une certaine proportion. Notre procédure d'estimation est basée sur des méthodes de sélection de modèle. Nous considérons une collection de modèles $\{S_m \times \Sigma_m, m \in \mathcal{M}\}$ où les S_m et les Σ_m sont des espaces linéaires. A chaque $m \in \mathcal{M}$, nous associons un couple d'estimateurs $(\hat{s}_m, \hat{\sigma}_m)$ de (s, σ) à valeurs dans $S_m \times \Sigma_m$. Ensuite, nous décrivons une procédure de sélection de modèle pour choisir un \hat{m} dans \mathcal{M} de telle sorte que le risque de Kullback de $(\hat{s}_{\hat{m}}, \hat{\sigma}_{\hat{m}})$ soit aussi proche que possible du minimum des risques de Kullback des estimateurs de la collection $\{(\hat{s}_m, \hat{\sigma}_m), m \in \mathcal{M}\}$. Des vitesses de convergence sur les boules de Hölder sont alors déduites pour la paire d'estimateurs $(\hat{s}_{\hat{m}}, \hat{\sigma}_{\hat{m}})$. Nous finissons par présenter des simulations pour illustrer les performances de nos estimateurs en pratique.

2.1. Introduction

Let us consider the statistical framework given by the distribution of a Gaussian vector Y with mean $s = (s_1, \dots, s_n)' \in \mathbb{R}^n$ and diagonal covariance matrix

$$\Gamma_\sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n \end{pmatrix}$$

where $\sigma = (\sigma_1, \dots, \sigma_n)' \in (0, \infty)^n$. The vectors s and σ are both assumed to be unknown. Hereafter, for any $t = (t_1, \dots, t_n)' \in \mathbb{R}^n$ and $\tau = (\tau_1, \dots, \tau_n)' \in (0, \infty)^n$, we denote by $P_{t,\tau}$ the distribution of a Gaussian vector with mean t and covariance matrix Γ_τ and by $\mathcal{K}(P_{s,\sigma}, P_{t,\tau})$ the *Kullback-Leibler divergence* between $P_{s,\sigma}$ and $P_{t,\tau}$,

$$\mathcal{K}(P_{s,\sigma}, P_{t,\tau}) = \frac{1}{2} \sum_{i=1}^n \frac{(s_i - t_i)^2}{\tau_i} + \phi\left(\frac{\tau_i}{\sigma_i}\right),$$

where $\phi(u) = \log u + 1/u - 1$, for $u > 0$. Note that, if the σ_i 's are known and constant, the Kullback-Leibler divergence becomes the squared L^2 -norm and, in expectation, corresponds to the quadratic risk.

Let us suppose that we observe two independent copies of Y , namely $Y^{[1]} = (Y_1^{[1]}, \dots, Y_n^{[1]})'$ and $Y^{[2]} = (Y_1^{[2]}, \dots, Y_n^{[2]})'$. Their coordinates can be expanded as

$$Y_i^{[j]} = s_i + \sqrt{\sigma_i} \varepsilon_i^{[j]}, \quad i = 1, \dots, n \text{ and } j = 1, 2, \quad (2.1.1)$$

where $\varepsilon^{[1]} = (\varepsilon_1^{[1]}, \dots, \varepsilon_n^{[1]})'$ and $\varepsilon^{[2]} = (\varepsilon_1^{[2]}, \dots, \varepsilon_n^{[2]})'$ are two independent standard Gaussian vectors. We are interested here in the estimation of the two vectors s and σ . Indeed, their behaviors contain substantial knowledge about the phenomenon represented by the distribution of Y . We have particularly in mind the case of a variance that stays approximately constant by periods and that can take several values in the proceeding of the observations. Of course, we want to estimate the mean s but, in this particular case, we are also interested in recovering the periods of constancy and the values taken by the variance σ . The Kullback-Leibler divergence measures the differences between two distributions $P_{s,\sigma}$ and $P_{t,\tau}$. Thus, it allows us to deal with the two estimation problems at the same time. More generally, the aim of this chapter is to estimate the pair (s, σ) by model selection on the basis of the observation of $Y^{[1]}$ and $Y^{[2]}$.

For this, we introduce a collection $\mathcal{F} = \{S_m \times \Sigma_m, m \in \mathcal{M}\}$ of products of linear subspaces of \mathbb{R}^n indexed by a finite or countable set \mathcal{M} . In the sequel, these products will be called *models* and, for any $m \in \mathcal{M}$, we will denote by D_m the dimension of $S_m \times \Sigma_m$. To each $m \in \mathcal{M}$, we will associate a pair of estimators $(\hat{s}_m, \hat{\sigma}_m)$ that is similar to the *maximum likelihood estimator* (MLE). It is well known that, if the σ_i 's are equal, the estimators of the mean and the variance factor given by maximization of the likelihood are independent. This fact does not remain true if the σ_i 's are not constant. To recover the independence between the estimators of the mean and the variance, we construct them separately from the two independent copies $Y^{[1]}$ and $Y^{[2]}$. For the estimator \hat{s}_m of s , we take the MLE based on $Y^{[1]}$ and for the estimator $\hat{\sigma}_m$ of σ , we take the MLE based on $Y^{[2]}$. Thus, for each $m \in \mathcal{M}$, we have a pair of independent estimators $(\hat{s}_m, \hat{\sigma}_m) = (\hat{s}_m(Y^{[1]}), \hat{\sigma}_m(Y^{[2]}))$ with values in $S_m \times \Sigma_m$. The *Kullback risk* of $(\hat{s}_m, \hat{\sigma}_m)$ is given by $\mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})]$ and is of order of the sum of two terms,

$$\inf_{(t,\tau) \in S_m \times \Sigma_m} \mathcal{K}(P_{s,\sigma}, P_{t,\tau}) + D_m. \quad (2.1.2)$$

The first one, called the *bias term*, represents the capacity of $S_m \times \Sigma_m$ to approximate the true value of (s, σ) . The second, called the *variance term*, is proportional to the dimension of the model and corresponds to the amount of noise that we have to control. To warrant a small risk, these two terms have to be small simultaneously. Indeed, using the Kullback risk as a quality criterion, a good model is one minimizing (2.1.2) among \mathcal{F} . Clearly, the choice of a such model depends on the pair of the unknown parameters (s, σ) and make good models unavailable to us. So, we have to construct a procedure to select an index $\hat{m} = \hat{m}(Y^{[1]}, Y^{[2]}) \in \mathcal{M}$ depending on the data only, such that $\mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_{\hat{m}}, \hat{\sigma}_{\hat{m}}})]$ is close to the smaller risk

$$R(s, \sigma, \mathcal{F}) = \inf_{m \in \mathcal{M}} \mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})].$$

The art of *model selection* is precisely to provide procedure solely based on the observations in that way. The classical way consists in minimizing an empirical penalized criterion stochastically close to the risk. Considering the *likelihood function* with respect to $Y^{[1]}$,

$$\forall t \in \mathbb{R}^n, \tau \in (0, \infty)^n, \mathcal{L}(t, \tau) = \frac{1}{2} \sum_{i=1}^n \frac{(Y_i^{[1]} - t_i)^2}{\tau_i} + \log \tau_i,$$

we choose \hat{m} as the minimizer over \mathcal{M} of the penalized likelihood criterion

$$\text{Crit}(m) = \mathcal{L}(\hat{s}_m, \hat{\sigma}_m) + \text{pen}(m) \quad (2.1.3)$$

where pen is a *penalty* function mapping \mathcal{M} into $\mathbb{R}_+ = [0, \infty)$. In this work, we give a form for the penalty in such a way to obtain a pair of estimators $(\hat{s}_{\hat{m}}, \hat{\sigma}_{\hat{m}})$ with a Kullback risk close to $R(s, \sigma, \mathcal{F})$.

Our approach is free of any prior assumption on s but requires that we know some upper bound $\gamma \geq 1$ on the ratio

$$\sigma^*/\sigma_* \leq \gamma$$

where σ^* (resp. σ_*) is the maximum (resp. minimum) of the σ_i 's. The knowledge of γ allows us to deal equivalently with two different cases. First, " $\gamma = 1$ " corresponds to the *homoscedastic* case where the components of $Y^{[1]}$ and $Y^{[2]}$ are independent with a common variance (*i.e.* $\sigma_i \equiv \sigma$) which can be unknown. On the other side, " $\gamma > 1$ " means that the σ_i 's can be distinct and are allowed to vary within some range. This uncommonness of the variances of the observations is known as the *heteroscedastic* case. Heteroscedasticity arises in many practical situations in which the assumption that the variances of the data are equal is debatable.

The research field of the model selection has known an important development in the last decades and it is beyond the scope of this chapter to make an exhaustive historical review of the domain. The interested reader could find a good introduction to model selection in the first chapters of [MT98]. The first heuristics in the domain are due to Mallows [Mal73] for the estimation of the mean in homoscedastic Gaussian regression with known variance. In more general Gaussian framework with common known variance, Barron *et al.* [BBM99], Birgé and Massart ([BM01a] and [BM01b]) have designed an adaptive model selection procedure to estimate the mean for quadratic risk. They provide non-asymptotic upper bound for the risk of the selected estimator. For bound of order of the smaller risk among the collection of models, this kind of result is called *oracle inequalities*. Baraud [Bar00] has generalized their results to homoscedastic statistical models with non-Gaussian noise admitting moment of order larger than 2 and a known variance. All these results remain true for common unknown variance if some upper bound on it is supposed to be known. Of course, the bigger is this bound, the worst are the results. Assuming that γ is known does not imply the knowledge of a such upper bound.

In the homoscedastic Gaussian framework with unknown variance, Akaike has proposed penalties for estimating the mean for quadratic risk (see [Aka70], [Aka73] and [Aka74]). Replacing the variance by a particular estimator in his penalty term, Baraud [Bar00] has obtained oracle inequalities for more general noise than Gaussian and polynomial collection of models. Recently, Baraud, Giraud and Huet [BGH09] have constructed penalties able to take into account the complexity of the collection of models for estimating the mean with quadratic risk in Gaussian homoscedastic model with unknown variance. They have also proved results for the estimation of the mean and the variance factor with Kullback risk. This problem is close to ours and corresponds to the case " $\gamma = 1$ ". A motivation for the present work was to extend their results to the heteroscedastic case " $\gamma > 1$ " in order to get oracle inequalities by minimization of penalized criterion as (2.1.3). Assuming that the collection of models is not too large, we obtain inequalities with the same flavor up to a logarithmic factor

$$\begin{aligned} & \mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] \\ & \leq C \inf_{m \in \mathcal{M}} \left\{ \inf_{(t,\tau) \in S_m \times \Sigma_m} \mathcal{K}(P_{s,\sigma}, P_{t,\tau}) + D_m \log^{1+\epsilon} D_m \right\} + R \end{aligned} \quad (2.1.4)$$

where C and R are positive constants depending in particular on γ and ϵ is a positive parameter.

A non-asymptotic model selection approach for estimation problem in heteroscedastic Gaussian model was studied in few papers only. In the chapter 6 of [Arl07], Arlot estimates the mean in heteroscedastic regression framework but for bounded data. For polynomial collection of models, he uses resampling penalties to get oracle inequalities for quadratic risk. Recently, Galtchouk and Pergamenschikov [GP05] have provided an adaptive nonparametric estimation procedure for the mean in a heteroscedastic Gaussian regression model. They obtain an oracle inequality for the quadratic risk under some regularity assumptions. Closer to our problem, Comte and Rozenholc [CR02] have estimated the pair (s, σ) . Their estimation procedure is different from ours and it makes the theoretical results difficultly comparable between us. For instance, they proceed in two steps (one for the mean and one for the variance) and they give risk bounds separately for each parameter in L_2 -norm while we estimate directly the pair (s, σ) for Kullback risk.

As described in [BM97], one of the main advantages of inequalities such as (2.1.4) is that they allow us to derive uniform convergence rates for the risk of the selected estimator over many classes of smoothness. Considering a collection of histogram models, we provide convergence rates over

Hölderian balls. Indeed, for $\alpha_1, \alpha_2 \in (0, 1]$, if s is α_1 -Hölderian and σ is α_2 -Hölderian, we prove that the risk of $(\hat{s}_m, \hat{\sigma}_m)$ converges with a rate of order of

$$\left(\frac{n}{\log^{1+\epsilon} n} \right)^{-2\alpha/(2\alpha+1)}$$

where $\alpha = \min\{\alpha_1, \alpha_2\}$ is the worst regularity. To compare this rate, we can think of the homoscedastic case with only one observation of Y . Indeed, in this case, the optimal rate of convergence in the minimax sense is $n^{-2\alpha/(2\alpha+1)}$ and, up to a logarithmic loss, our rate is comparable to this one. To our knowledge, our results in non-asymptotic estimation of the mean and the variance in heteroscedastic Gaussian model are new.

The chapter is organized as follows. The main results are presented in section 2.2. In section 2.3, we carry out a simulation study in order to illustrate the performances of our estimators in practice with the Kullback risk and the quadratic risk. The last sections are devoted to the proofs and to some technical results.

2.2. Main results

In a first time, we introduce the collection of models, the estimators and the procedure. Next, we present the main results whose proofs can be found in the section 2.4. In the sequel, we consider the framework (2.1.1) and, for the sake of simplicity, we suppose that there exists an integer $k_n \geq 0$ such that $n = 2^{k_n}$.

2.2.1. Model collection and estimators. In order to estimate the mean and the variance, we consider linear subspaces of \mathbb{R}^n constructed as follows. Let \mathcal{M} be a countable or finite set. To each $m \in \mathcal{M}$, we associate a regular partition p_m of $\{1, \dots, 2^{k_n}\}$ given by the $|p_m| = 2^{k_m}$ consecutive blocks

$$\{(i-1)2^{k_n-k_m} + 1, \dots, i2^{k_n-k_m}\}, \quad i = 1, \dots, |p_m|.$$

For any $I \in p_m$ and any $x \in \mathbb{R}^n$, let us denote by $x|_I$ the vector of $\mathbb{R}^{|p_m|}$ with coordinates $(x_i)_{i \in I}$. Then, to each $m \in \mathcal{M}$, we also associate a linear subspace E_m of $\mathbb{R}^{|p_m|}$ with dimension $1 \leq d_m \leq 2^{k_n-k_m}$. This set of pairs (p_m, E_m) allows us to construct a collection of models. Hereafter, we identify each $m \in \mathcal{M}$ to its corresponding pair (p_m, E_m) .

For any $m = (p_m, E_m) \in \mathcal{M}$, we introduce the subspace $S_m \subset \mathbb{R}^n$ of the E_m -piecewise vectors,

$$S_m = \{x \in \mathbb{R}^n \text{ such that } \forall I \in p_m, x|_I \in E_m\},$$

and the subspace $\Sigma_m \subset \mathbb{R}^n$ of the piecewise constant vectors,

$$\Sigma_m = \left\{ \sum_{I \in p_m} g_I \mathbb{1}_I, \quad \forall I \in p_m, g_I \in \mathbb{R} \right\}.$$

The dimension of $S_m \times \Sigma_m$ is denoted by $D_m = |p_m|(d_m + 1)$. To estimate the pair (s, σ) , we only deal with models $S_m \times \Sigma_m$ constructed in a such way. More precisely, we consider a collection of products of linear subspaces

$$\mathcal{F} = \{S_m \times \Sigma_m, m \in \mathcal{M}\} \tag{2.2.1}$$

where \mathcal{M} is a set of pairs (p_m, E_m) as above. In the chapter, we will often make the following hypothesis on the collection of models:

(H_θ): There exists $\theta > 1$ such that

$$\forall m \in \mathcal{M}, n \geq \frac{\theta}{\theta-1}(\gamma+2)D_m.$$

This hypothesis avoids handling models with dimension too great with respect to the number of observations.

Let $m \in \mathcal{M}$, we denote by π_m the orthogonal projection on S_m . We estimate (s, σ) by the pair of independent estimators $(\hat{s}_m, \hat{\sigma}_m) \in S_m \times \Sigma_m$ given by

$$\hat{s}_m = \pi_m Y^{[1]}$$

and

$$\hat{\sigma}_m = \sum_{I \in p_m} \hat{\sigma}_{m,I} \mathbb{1}_I \text{ where } \forall I \in p_m, \hat{\sigma}_{m,I} = \frac{1}{|I|} \sum_{i \in I} \left(Y_i^{[2]} - \left(\pi_m Y^{[2]} \right)_i \right)^2.$$

Thus, we get a collection of estimators $\{(\hat{s}_m, \hat{\sigma}_m), m \in \mathcal{M}\}$.

2.2.2. Risk upper bound. We first study the risk on a single model to understand its order. Take an arbitrary $m \in \mathcal{M}$. We define $(s_m, \sigma_m) \in S_m \times \Sigma_m$ by

$$s_m = \pi_m s$$

and

$$\sigma_m = \sum_{I \in p_m} \sigma_{m,I} \mathbb{1}_I \text{ where } \forall I \in p_m, \sigma_{m,I} = \frac{1}{|I|} \sum_{i \in I} (s_i - s_{m,i})^2 + \sigma_i.$$

Easy computations proves that the pair (s_m, σ_m) reaches the minimum of the Kullback-Leibler divergence on $S_m \times \Sigma_m$,

$$\begin{aligned} \inf_{(t, \tau) \in S_m \times \Sigma_m} \mathcal{K}(P_{s, \sigma}, P_{t, \tau}) &= \mathcal{K}(P_{s, \sigma}, P_{s_m, \sigma_m}) \\ &= \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \log \left(\frac{\sigma_{m,I}}{\sigma_i} \right). \end{aligned} \quad (2.2.2)$$

The next proposition allows us to compare this quantity with the Kullback risk of $(\hat{s}_m, \hat{\sigma}_m)$.

PROPOSITION 2.1. *Let $m \in \mathcal{M}$, if the hypothesis (\mathbf{H}_θ) is fulfilled, then*

$$\mathcal{K}(P_{s, \sigma}, P_{s_m, \sigma_m}) \vee \frac{D_m}{4\gamma} \leq \mathbb{E}[\mathcal{K}(P_{s, \sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] \leq \mathcal{K}(P_{s, \sigma}, P_{s_m, \sigma_m}) + \kappa \gamma^2 \theta^2 D_m$$

where $\kappa > 1$ is a constant that can be taken equal to $1 + 2e^{-1}$.

As announced in (2.1.2), this result shows that the Kullback risk of the pair $(\hat{s}_m, \hat{\sigma}_m)$ is of order of the sum of a bias term $\mathcal{K}(P_{s, \sigma}, P_{s_m, \sigma_m})$ and a variance term which is proportional to D_m . Thus, minimizing the risk $\mathbb{E}[\mathcal{K}(P_{s, \sigma}, P_{\hat{s}_m, \hat{\sigma}_m})]$ among $m \in \mathcal{M}$ corresponds to finding a model that realizes a trade-off between these two terms.

Let pen be a non negative function on \mathcal{M} , $\hat{m} \in \mathcal{M}$ is any minimizer of the penalized criterion

$$\hat{m} \in \underset{m \in \mathcal{M}}{\text{argmin}} \{ \mathcal{L}(\hat{s}_m, \hat{\sigma}_m) + \text{pen}(m) \}. \quad (2.2.3)$$

In the sequel, we denote by $(\tilde{s}, \tilde{\sigma}) = (\hat{s}_{\hat{m}}, \hat{\sigma}_{\hat{m}})$ the selected pair of estimators. It satisfies the following result:

THEOREM 2.2. *Under the hypothesis (\mathbf{H}_θ) , suppose there exist $A, B > 0$ such that, for any $(k, d) \in \mathbb{N}^2$,*

$$M_{k,d} = \text{Card} \{ m \in \mathcal{M} \text{ such that } |p_m| = 2^k \text{ and } d_m = d \} \leq A(1+d)^B \quad (2.2.4)$$

where \mathcal{M} is the set defined at the beginning of the section 2.2.1. Moreover, assume that there exist $\delta, \epsilon > 0$ such that

$$D_m \leq \frac{5\delta\gamma n}{\log^{1+\epsilon} n}, \quad \forall m \in \mathcal{M}. \quad (2.2.5)$$

If we take

$$\forall m \in \mathcal{M}, \text{pen}(m) = (\gamma\theta + \log^{1+\epsilon} D_m) D_m \quad (2.2.6)$$

then

$$\mathbb{E}[\mathcal{K}(P_{s, \sigma}, P_{\tilde{s}, \tilde{\sigma}})] \leq C \inf_{m \in \mathcal{M}} \{ \mathcal{K}(P_{s, \sigma}, P_{s_m, \sigma_m}) + D_m \log^{1+\epsilon} D_m \} + R \quad (2.2.7)$$

where $R = R(\gamma, \theta, A, B, \epsilon, \delta)$ is a positive constant and C can be taken equal to

$$C = 2 \left(1 + \frac{(\kappa\gamma\theta + 1)\gamma\theta}{\log^{1+\epsilon} 2} \right).$$

The inequality (2.2.7) is close to an oracle inequality up to a logarithmic factor. Thus, considering the penalty (2.2.6) whose order is slightly larger than the dimension of the model, the risk of the estimator provided by the criterion (2.1.3) is comparable to the minimum among the collection of models \mathcal{F} .

2.2.3. Convergence rate. One of the main advantages of an inequality as (2.2.7) is that it gives uniform convergence rates with respect to many well known classes of smoothness. To illustrate this, we consider the particular case of the regression on a fixed design. For example, in the framework (2.1.1), we suppose that

$$\forall 1 \leq i \leq n, \quad s_i = s_r(i/n) \text{ and } \sigma_i = \sigma_r(i/n),$$

where s_r and σ_r are two unknown functions that map $[0, 1]$ to \mathbb{R} .

In this section, we handle the normalized Kullback-Leibler divergence

$$\mathcal{K}_n(P_{s,\sigma}, P_{t,\tau}) = \frac{1}{n} \mathcal{K}(P_{s,\sigma}, P_{t,\tau}) ,$$

and, for any $\alpha \in (0, 1)$ and any $L > 0$, we denote by $\mathcal{H}_\alpha(L)$ the space of the α -Hölderian functions with constant L on $[0, 1]$,

$$\mathcal{H}_\alpha(L) = \{f : [0, 1] \rightarrow \mathbb{R} : \forall x, y \in [0, 1], |f(x) - f(y)| \leq L|x - y|^\alpha\} .$$

Moreover, we consider a collection of models \mathcal{F}^{PC} as described in the section 2.2.1 such that, for any $m \in \mathcal{M}$, E_m is the space of dyadic piecewise constant functions on d_m blocks. More precisely, let $m = (p_m, E_m) \in \mathcal{M}$ and consider the regular dyadic partition p'_m with $|p_m|d_m$ blocks that is a refinement of p_m . We define S_m as the space of the piecewise constant functions on p'_m ,

$$S_m = \left\{ f = \sum_{I \in p'_m} f_I \mathbb{1}_I \text{ such that } \forall I \in p'_m, f_I \in \mathbb{R} \right\} ,$$

and Σ_m as the space of the piecewise constant functions on p_m ,

$$\Sigma_m = \left\{ g = \sum_{I \in p_m} g_I \mathbb{1}_I \text{ such that } \forall I \in p_m, g_I \in \mathbb{R} \right\} .$$

Then, the collection of models that we consider is

$$\mathcal{F}^{PC} = \{S_m \times \Sigma_m, m \in \mathcal{M}\} .$$

Note that this collection satisfies (2.2.4) with $A = 1$ and $B = 0$. The following result gives a uniform convergence rate for $(\tilde{s}, \tilde{\sigma})$ over Hölderian balls.

PROPOSITION 2.3. *Let $\alpha_1, \alpha_2 \in (0, 1]$, $L_1, L_2 > 0$ and assume that (\mathbf{H}_θ) is fulfilled. Consider the collection of models \mathcal{F}^{PC} and $\delta, \epsilon > 0$ such that, for any $m \in \mathcal{M}$,*

$$D_m \leq \frac{5\delta\gamma n}{\log^{1+\epsilon} n} .$$

Denoting by $(\tilde{s}, \tilde{\sigma})$ the estimator selected via the penalty (2.2.6), if n satisfies

$$n \geq \left(\frac{2\sigma_*^2}{L_1^2 \sigma_* + L_2^2} \right)^2 \vee e^{4(1+\epsilon)^2}$$

then

$$\sup_{(s_r, \sigma_r) \in \mathcal{H}_{\alpha_1}(L_1) \times \mathcal{H}_{\alpha_2}(L_2)} \mathbb{E} [\mathcal{K}_n(P_{s,\sigma}, P_{\tilde{s}, \tilde{\sigma}})] \leq C \left(\frac{n}{\log^{1+\epsilon} n} \right)^{-2\alpha/(2\alpha+1)} \quad (2.2.8)$$

where $\alpha = \min\{\alpha_1, \alpha_2\}$ and C is a constant which depends on $\alpha_1, \alpha_2, L_1, L_2, \theta, \gamma, \sigma_, \delta$ and ϵ .*

For the estimation of the mean s in quadratic risk with one observation of Y , Galtchouk and Pergamenschikov [GP05] have computed the heteroscedastic minimax risk. Under some assumptions on the regularity of σ_r and assuming that $s_r \in \mathcal{H}_{\alpha_1}(L_1)$, they show that the order of the optimal rate of convergence in minimax sense is $C_{\alpha_1, \sigma} n^{-2\alpha_1/(2\alpha_1+1)}$. Concerning the estimation of the variance vector σ in quadratic risk with one observation of Y and unknown mean, Wang *et al.* [Wt08] have proved that the order of the minimax rate of convergence for the estimation of σ is $C_{\alpha_1, \alpha_2} \max\{n^{-4\alpha_1}, n^{-2\alpha_2/(2\alpha_2+1)}\}$ once $s_r \in \mathcal{H}_{\alpha_1}(L_1)$ and $\sigma_r \in \mathcal{H}_{\alpha_2}(L_2)$. For $\alpha_1, \alpha_2 \in (0, 1]$ the maximum of these two rates is of order $n^{-2\alpha/(2\alpha+1)}$ where $\alpha = \min\{\alpha_1, \alpha_2\}$ is the worst among the regularities of s_r and σ_r . Up to a logarithmic term, the rate of convergence over Hölderian balls given by our procedure recover this rate for the Kullback risk.

2.3. Simulation study

To illustrate our results, we consider the following pairs of functions (s_r, σ_r) defined on $[0, 1]$ and, for each one, we precise the true value of γ :

- M1 ($\gamma = 2$)

$$s_r(x) = \begin{cases} 4 & \text{if } 0 \leq x < 1/4 \\ 0 & \text{if } 1/4 \leq x < 1/2 \\ 2 & \text{if } 1/2 \leq x < 3/4 \\ 1 & \text{if } 3/4 \leq x \leq 1 \end{cases} \quad \text{and} \quad \sigma_r(x) = \begin{cases} 2 & \text{if } 0 \leq x < 1/2 \\ 1 & \text{if } 1/2 \leq x \leq 1 \end{cases} ,$$

- M2 ($\gamma = 1$)

$$s_r(x) = 1 + \sin(2\pi x + \pi/3) \quad \text{and} \quad \sigma_r(x) = 1 ,$$

- M3 ($\gamma = 7/3$)

$$s_r(x) = 3x/2 \quad \text{and} \quad \sigma_r(x) = 1/2 + 2\sin(4\pi(x \wedge 1/2)^2)/3 ,$$

- M4 ($\gamma = 2$)

$$s_r(x) = 1 + \sin(4\pi(x \wedge 1/2)) \quad \text{and} \quad \sigma_r(x) = (3 + \sin(2\pi x))/2 .$$

In all this section, we consider the collection of models \mathcal{F}^{PC} and we take $n = 1024$ (*i.e.* $k_n = 10$). Let us first present how our procedure performs on the examples with the true value of γ for each simulation, $\epsilon = 10^{-2}$ and $\delta = 3$ in the assumption (2.2.5) and the penalty (2.2.6) with $\theta = 2$. The estimators are drawn in plain line and the true functions in dotted line. In the case of M1, we can note that the procedure choose the “good” model in the sense that if the pair (s_r, σ_r) belongs to a model of \mathcal{F}^{PC} , this one is generally chosen by our procedure. Repeating the simulation 100 000

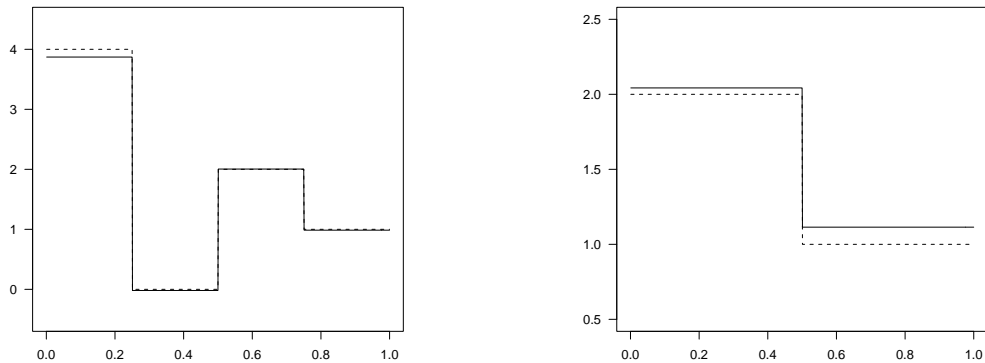


FIGURE 2.1. Estimation on the mean (left) and the variance (right) in the case M1.

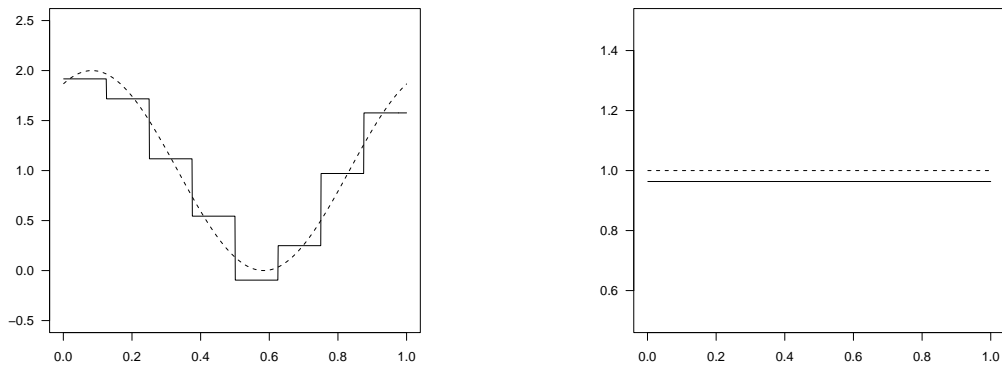


FIGURE 2.2. Estimation on the mean (left) and the variance (right) in the case M2.

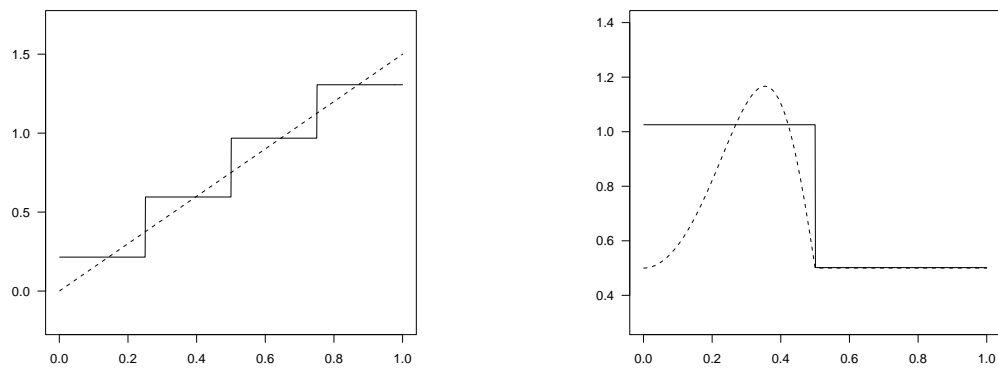


FIGURE 2.3. Estimation on the mean (left) and the variance (right) in the case M3.

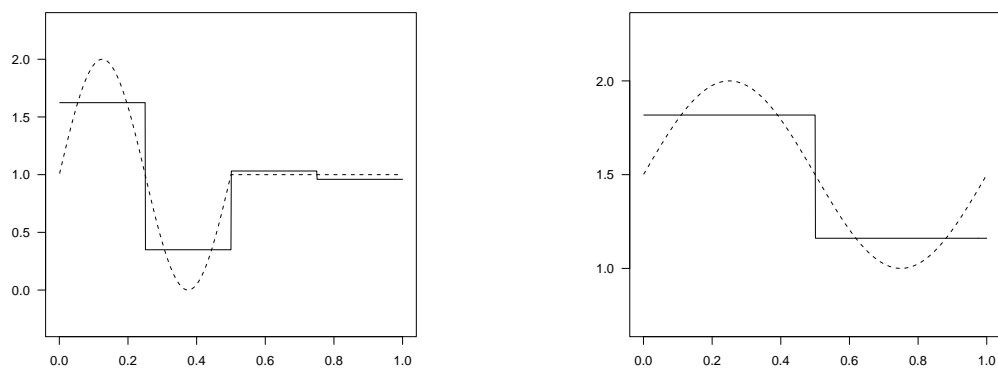


FIGURE 2.4. Estimation on the mean (left) and the variance (right) in the case M4.

times with the framework of M1 gives us that, with probability higher than 99.9%, the probability for making this “good” choice is about $0.9978 (\pm 4 \times 10^{-4})$. Even if the mean does not belong to one of the S_m ’s, the procedure recover the homoscedastic nature of the observations in the case M2. By doing 100 000 simulations with the framework induced by M2, the probability to choose an homoscedastic model is around $0.99996 (\pm 1 \times 10^{-5})$ with a confidence of 99.9%. For more general framework as M3 and M4, the estimators perform visually well and detect the changements in the behaviour of the mean and the variance functions.

The parameter γ is supposed to be known and is present in the definition of the penalty (2.2.6). So, we naturally can ask what is its importance in the procedure. In particular, what happens if we do not have the good value? The following table present some estimations of the ratio

$$\mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\tilde{s},\tilde{\sigma}})] / \inf_{m \in \mathcal{M}} \mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})]$$

for several values of γ . These estimated values have been obtained with 500 repetitions for each one. The main part of the computation time is devoted to the estimation of the oracle’s risk. In the cases

γ	1	1.5	2	2.5	3
M1	0.98	1.02	1.02	1.04	1.01
M2	1.49	1.59	1.88	2.29	2.89
M3	1.77	1.78	1.81	1.90	1.94
M4	1.25	1.26	1.27	1.32	1.33

TABLE 2.1. Ratio between the Kullback risk of $(\tilde{s}, \tilde{\sigma})$ and the one of the oracle.

M1, M3 and M4, the ratio does not suffer to much from small errors on the knowledge of γ . The more affected case is the homoscedastic one but we see that the best estimation is obtained for the good value of γ as we could expect. More generally, it is interesting to observe that, even if there is a small error on the value of γ , the ratio stays reasonably small.

In the regression framework with heteroscedastic noise, we can be interested in separate estimations of the mean and the variance functions. Because our procedure provide a simultaneous estimation of these two functions, we can ask how perform our estimators \tilde{s} and $\tilde{\sigma}$ individually. Considering the quadratic risks $\mathbb{E}[\|s - \tilde{s}\|^2]$ and $\mathbb{E}[\|\sigma - \tilde{\sigma}\|^2]$ of \tilde{s} and $\tilde{\sigma}$ respectively, it could be interesting to compare them to the minimal quadratic risk among the collection of estimators. To illustrate this, we give below two sets of estimations of the ratios

$$\mathbb{E}[\|s - \tilde{s}\|^2] / \inf_{m \in \mathcal{M}} \mathbb{E}[\|s - \hat{s}_m\|^2] \quad \text{and} \quad \mathbb{E}[\|\sigma - \tilde{\sigma}\|^2] / \inf_{m \in \mathcal{M}} \mathbb{E}[\|\sigma - \hat{\sigma}_m\|^2]$$

in the frameworks presented in the beginning of this section. We can observe on the following estimations that the quadratic risks of our estimators are quite close to the minimal ones among the collection of models.

γ	1	1.5	2	2.5	3
M1	0.98	1.01	0.95	1.04	0.98
M2	1.52	1.67	2.04	2.43	3.04
M3	1.73	1.76	1.82	1.88	1.96
M4	1.47	1.48	1.47	1.47	1.49

TABLE 2.2. Ratio between the L_2 -risk of \tilde{s} and the minimal one among the \hat{s}_m ’s.

γ	1	1.5	2	2.5	3
M1	1.00	1.06	1.03	1.02	1.01
M2	1.11	1.56	1.68	2.21	3.36
M3	2.02	2.07	2.13	2.20	2.23
M4	1.18	1.37	1.34	1.44	1.49

TABLE 2.3. Ratio between the L_2 -risk of $\tilde{\sigma}$ and the minimal one among the $\hat{\sigma}_m$'s.

2.4. Proofs

For any $I \subset \{1, \dots, n\}$ and any $x, y \in \mathbb{R}^n$, we introduce the notations

$$\langle x, y \rangle_I = \sum_{i \in I} x_i y_i \quad \text{and} \quad \|x\|_I^2 = \sum_{i \in I} x_i^2 .$$

Let $m \in \mathcal{M}$, we will use several times in the proofs the fact that, for any $I \in p_m$,

$$|I| \hat{\sigma}_{m,I} \geq \sigma_* \chi^2(|I| - d_m - 1) \quad (2.4.1)$$

where $\chi^2(|I| - d_m - 1)$ is a χ^2 random variable with $|I| - d_m - 1$ degrees of freedom.

2.4.1. Proof of Proposition 2.1. Recalling (2.2.2) and using the independence between \hat{s}_m and $\hat{\sigma}_m$, we expand the Kullback risk of $(\hat{s}_m, \hat{\sigma}_m)$,

$$\begin{aligned} \mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] &= \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \mathbb{E} \left[\frac{(s_i - \hat{s}_{m,i})^2}{\hat{\sigma}_{m,I}} + \phi \left(\frac{\hat{\sigma}_{m,I}}{\sigma_i} \right) \right] \\ &= \frac{1}{2} \sum_{I \in p_m} \mathbb{E} \left[\frac{1}{\hat{\sigma}_{m,I}} \right] \mathbb{E} [\|s - \hat{s}_m\|_I^2] \\ &\quad + \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \mathbb{E} \left[\log \frac{\hat{\sigma}_{m,I}}{\sigma_{m,I}} + \frac{\sigma_i}{\hat{\sigma}_{m,I}} - 1 \right] + \log \frac{\sigma_{m,I}}{\sigma_i} \\ &= \mathcal{K}(P_{s,\sigma}, P_{s_m, \sigma_m}) + \frac{1}{2} \sum_{I \in p_m} |I| \mathbb{E} \left[\phi \left(\frac{\hat{\sigma}_{m,I}}{\sigma_{m,I}} \right) \right] \\ &\quad + \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \mathbb{E} \left[\frac{\sigma_i + (s_i - s_{m,i})^2 - \sigma_{m,I}}{\hat{\sigma}_{m,I}} \right] \\ &\quad + \frac{1}{2} \sum_{I \in p_m} \mathbb{E} \left[\frac{1}{\hat{\sigma}_{m,I}} \right] \mathbb{E} [\|\pi_m \Gamma_\sigma^{1/2} \varepsilon^{[1]}\|_I^2] \\ &= \mathcal{K}(P_{s,\sigma}, P_{s_m, \sigma_m}) + \mathbb{E}_1 + \mathbb{E}_2 \end{aligned} \quad (2.4.2)$$

where

$$\mathbb{E}_1 = \frac{1}{2} \sum_{I \in p_m} |I| \mathbb{E} \left[\phi \left(\frac{\hat{\sigma}_{m,I}}{\sigma_{m,I}} \right) \right] \quad \text{and} \quad \mathbb{E}_2 = \frac{1}{2} \sum_{I \in p_m} \mathbb{E} \left[\frac{1}{\hat{\sigma}_{m,I}} \right] \sum_{i \in I} \pi_{m,i,i} \sigma_i .$$

To upper bound the first expectation, note that

$$\forall I \in p_m, \quad \mathbb{E}[\hat{\sigma}_{m,I}] = \sigma_{m,I} - \frac{1}{|I|} \sum_{i \in I} \pi_{m,i,i} \sigma_i = \sigma_{m,I} (1 - \rho_I)$$

where

$$\rho_I = \frac{1}{|I| \sigma_{m,I}} \sum_{i \in I} \pi_{m,i,i} \sigma_i \in (0, 1) .$$

We apply the lemmas 2.10 and 2.11 to each block $I \in p_m$ and, by concavity of the logarithm, we get

$$\begin{aligned}
\mathbb{E} \left[\phi \left(\frac{\hat{\sigma}_{m,I}}{\sigma_{m,I}} \right) \right] &\leq \log \mathbb{E} \left[\frac{\hat{\sigma}_{m,I}}{\sigma_{m,I}} \right] + \mathbb{E} \left[\frac{\sigma_{m,I}}{\hat{\sigma}_{m,I}} \right] - 1 \\
&\leq \log(1 - \rho_I) + \frac{1}{1 - \rho_I} \left(1 + \frac{2\kappa\gamma^2}{|I| - d_m - 2} \right) - 1 \\
&\leq -\rho_I + \frac{1}{1 - \rho_I} \left(1 + \frac{2\kappa\gamma^2}{|I| - d_m - 2} \right) - 1 \\
&\leq \frac{1}{1 - \rho_I} \left(\rho_I^2 + \frac{2\kappa\gamma^2}{|I| - d_m - 2} \right) .
\end{aligned}$$

Using (\mathbf{H}_θ) and the fact that $\rho_I \leq \gamma d_m / |I|$, we obtain

$$\begin{aligned}
\mathbb{E}_1 &\leq \frac{1}{2} \sum_{I \in p_m} \frac{|I|}{1 - \rho_I} \left(\rho_I^2 + \frac{2\kappa\gamma^2}{|I| - d_m - 2} \right) \\
&\leq \frac{1}{2} \sum_{I \in p_m} \frac{\gamma^2 d_m^2}{|I| - \gamma d_m} + \frac{2\kappa\gamma^2 |I|^2}{(|I| - \gamma d_m)(|I| - d_m - 2)} \\
&\leq \frac{\gamma^2 \theta |p_m| d_m}{2} + \kappa \gamma^2 \theta^2 |p_m| .
\end{aligned} \tag{2.4.4}$$

The second expectation in (2.4.3) is easier to upper bound by using (2.4.1) and the fact that $d_m \geq 1$,

$$\begin{aligned}
\mathbb{E}_2 &= \frac{1}{2} \sum_{I \in p_m} \mathbb{E} \left[\frac{1}{\hat{\sigma}_{m,I}} \right] \sum_{i \in I} \pi_{m,i,i} \sigma_i \\
&\leq \frac{1}{2} \sum_{I \in p_m} \frac{\gamma |I| d_m}{|I| - d_m - 3} \\
&\leq \frac{\gamma \theta |p_m| d_m}{2} .
\end{aligned} \tag{2.4.5}$$

We now sum (2.4.4) and (2.4.5) to obtain

$$\mathbb{E}_1 + \mathbb{E}_2 \leq \gamma^2 \theta |p_m| d_m + \kappa \gamma^2 \theta^2 |p_m| \leq \kappa \gamma^2 \theta^2 D_m .$$

For the lower bound, the positivity of ϕ in (2.4.2) and the independence between \hat{s}_m and $\hat{\sigma}_m$ give us

$$\begin{aligned}
\mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] &\geq \frac{1}{2} \sum_{I \in p_m} \mathbb{E} \left[\frac{\|s - \hat{s}_m\|_I^2}{\hat{\sigma}_{m,I}} \right] \\
&\geq \frac{1}{2} \sum_{I \in p_m} \frac{\mathbb{E} [\|s - \hat{s}_m\|_I^2]}{\mathbb{E} [\hat{\sigma}_{m,I}]} \\
&\geq \frac{1}{2} \sum_{I \in p_m} |I| \frac{\|s - s_m\|_I^2 + \sigma_* d_m}{\|s - s_m\|_I^2 + (|I| - d_m) \sigma_*} .
\end{aligned}$$

It is obvious that the hypothesis (\mathbf{H}_θ) ensures $d_m \leq |I|/2$. Thus, we get $\sigma_* d_m \leq (|I| - d_m) \sigma_*$ and

$$\mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] \geq \frac{1}{2} \sum_{I \in p_m} \frac{|I| \sigma_* d_m}{(|I| - d_m) \sigma_*} \geq \frac{|p_m| d_m}{2\gamma} \geq \frac{D_m}{4\gamma} .$$

To conclude, we know that $(\hat{s}_m, \hat{\sigma}_m) \in S_m \times \Sigma_m$ and, by definition of (s_m, σ_m) , it implies

$$\mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] \geq \mathcal{K}(P_{s,\sigma}, P_{s_m, \sigma_m}) .$$

2.4.2. Proof of Theorem 2.2. We prove the following more general result:

THEOREM 2.4. *Let $\alpha \in (0, 1)$ and consider a collection of positive weights $\{x_m\}_{m \in \mathcal{M}}$. If the hypothesis (\mathbf{H}_θ) is fulfilled and if*

$$\forall m \in \mathcal{M}, \text{pen}(m) \geq \gamma \theta D_m + x_m, \quad (2.4.6)$$

then

$$\begin{aligned} (1 - \alpha) \mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s},\hat{\sigma}})] \\ \leq \inf_{m \in \mathcal{M}} \{ \mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] + \text{pen}(m) \} + R_1(\mathcal{M}) + R_2(\mathcal{M}) \end{aligned}$$

where $R_1(\mathcal{M})$ and $R_2(\mathcal{M})$ are defined by

$$R_1(\mathcal{M}) = C \theta^2 \gamma \sum_{m \in \mathcal{M}} \sqrt{|p_m|} d_m \left(\frac{2C \theta^2 \gamma \sqrt{|p_m|} d_m \log(1 + d_m)}{x_m} \right)^{[2 \log(1 + d_m)]}$$

and

$$R_2(\mathcal{M}) = \frac{2(\alpha + \gamma \theta) + 1}{\alpha} \sum_{m \in \mathcal{M}} |p_m| \exp \left(- \frac{n}{2\theta |p_m|} \log \left(1 + \frac{\alpha |p_m| x_m}{\gamma n (\alpha + 2)} \right) \right).$$

In these expressions, $[\cdot]$ is the integral part and C is a positive constant that could be taken equal to $12\sqrt{2e}/(\sqrt{e} - 1)$.

Before proving this result, let us see how it implies the theorem 2.2. The choice (2.2.6) for the penalty function corresponds to $x_m = D_m \log^{1+\epsilon} D_m$ in (2.4.6). Applying the previous theorem with $\alpha = 1/2$ leads us to

$$\begin{aligned} \mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s},\hat{\sigma}})] \\ \leq 2 \inf_{m \in \mathcal{M}} \{ \mathbb{E} [\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] + \text{pen}(m) \} + 2C \theta^2 \gamma R_1 + 8(\gamma \theta + 1) R_2 \end{aligned}$$

with

$$R_1 = \sum_{m \in \mathcal{M}} \sqrt{|p_m|} d_m \left(\frac{2C \theta^2 \gamma \sqrt{|p_m|} d_m \log(1 + d_m)}{x_m} \right)^{[2 \log(1 + d_m)]}$$

and

$$R_2 = \sum_{m \in \mathcal{M}} |p_m| \exp \left(- \frac{n}{2\theta |p_m|} \log \left(1 + \frac{|p_m| x_m}{5\gamma n} \right) \right).$$

Using the upper bound on the risk of the proposition 2.1, we easily obtain the coefficient of the infimum in (2.2.7). Thus, it remains to prove that the two quantities R_1 and R_2 can be upper bounded independently of n . For this, we denote by $B' = B + 2 \log(2C \theta^2 \gamma) + 1$ and we compute

$$\begin{aligned} R_1 &= \sum_{m \in \mathcal{M}} \sqrt{|p_m|} d_m \left(\frac{2C \theta^2 \gamma \sqrt{|p_m|} d_m \log(1 + d_m)}{|p_m|(1 + d_m) \log^{1+\epsilon}(|p_m|(1 + d_m))} \right)^{[2 \log(1 + d_m)]} \\ &\leq \sum_{k \geq 0} \sum_{d \geq 1} M_{k,d} 2^{k/2} d \left(\frac{2C \theta^2 \gamma 2^{-k/2} \log(1 + d)}{(k \log 2 + \log(1 + d))^{1+\epsilon}} \right)^{[2 \log(1 + d)]} \\ &\leq A \sum_{k \geq 0} \sum_{d \geq 1} (1 + d)^{B'} 2^{k/2} \left(\frac{2^{-k/2} \log(1 + d)}{(k \log 2 + \log(1 + d))^{1+\epsilon}} \right)^{[2 \log(1 + d)]} \\ &\leq A(R'_1 + R''_1). \end{aligned}$$

We have split the sum in two terms, the first one is for $d = 1$,

$$R'_1 = \sum_{k \geq 0} \frac{2^{B'} \log 2}{(k \log 2 + \log 2)^{1+\epsilon}} = \frac{2^{B'}}{\log^\epsilon 2} \sum_{k \geq 0} \frac{1}{(k + 1)^{1+\epsilon}} < \infty.$$

The other part R_1'' is for $d \geq 2$ and is equal to

$$\sum_{k \geq 0} \sum_{d \geq 2} (1+d)^{B'} 2^{-k(\lfloor 2 \log(1+d) \rfloor - 1)/2} \left(\frac{\log(1+d)}{(k \log 2 + \log(1+d))^{1+\epsilon}} \right)^{\lfloor 2 \log(1+d) \rfloor}.$$

Noting that $1 < \log(1+d) \leq \lfloor 2 \log(1+d) \rfloor$, we have

$$\begin{aligned} R_1'' &\leq \sum_{k \geq 0} 2^{-k/2} \sum_{d \geq 2} (1+d)^{B'} \exp(-\epsilon \lfloor 2 \log(1+d) \rfloor \log \log(1+d)) \\ &\leq \frac{\sqrt{2}}{\sqrt{2}-1} \sum_{d \geq 2} (1+d)^{B' - \epsilon \log \log(1+d)} < \infty. \end{aligned}$$

We now handle R_2 . Our choice of $x_m = D_m \log^{1+\epsilon} D_m$ and the hypothesis (2.2.5) imply

$$\frac{|p_m| x_m}{5\gamma n} \leq \delta |p_m| = \frac{1 - (\delta |p_m| + 1)^{-1}}{(\delta |p_m| + 1)^{-1}}.$$

We recall that, for any $a \in (0, 1)$, if $0 \leq t \leq (1-a)/a$, then $\log(1+t) \geq at$. Take $a = (\delta |p_m| + 1)^{-1}$ to obtain

$$\log \left(1 + \frac{|p_m| x_m}{5\gamma n} \right) \geq \frac{|p_m| x_m}{5(\delta |p_m| + 1)\gamma n} \geq \frac{x_m}{5(\delta + 1)\gamma n}.$$

For any positive t , $1 + t^{1+\epsilon} \leq (1+t)^{1+\epsilon}$, then we finally obtain

$$\begin{aligned} R_2 &= \sum_{m \in \mathcal{M}} |p_m| \exp \left(-\frac{n}{2\theta |p_m|} \log \left(1 + \frac{|p_m| x_m}{5\gamma n} \right) \right) \\ &\leq \sum_{m \in \mathcal{M}} |p_m| \exp \left(-\frac{x_m}{10\theta \gamma (\delta + 1) |p_m|} \right) \\ &\leq \sum_{k \geq 0} \sum_{d \geq 1} M_{k,d} 2^k \exp \left(-\frac{(1+d) \log^{1+\epsilon} (2^k (1+d))}{10\theta \gamma (\delta + 1)} \right) \\ &\leq AR_2' R_2'' \end{aligned}$$

where we have set

$$R_2' = \sum_{k \geq 0} \exp \left(k \log 2 - \frac{(k \log 2)^{1+\epsilon}}{5\theta \gamma (\delta + 1)} \right) < \infty$$

and

$$R_2'' = \sum_{d \geq 1} \exp \left(B \log(1+d) - \frac{(1+d) \log^{1+\epsilon} (1+d)}{10\theta \gamma (\delta + 1)} \right) < \infty.$$

We now have to prove theorem 2.4. For an arbitrary $m \in \mathcal{M}$, we begin the proof by expanding the Kullback-Leibler divergence of $(\tilde{s}, \tilde{\sigma})$,

$$\begin{aligned} \mathcal{K}(P_{s,\sigma}, P_{\tilde{s},\tilde{\sigma}}) &= \frac{1}{2} \sum_{i=1}^n \frac{(s_i - \tilde{s}_i)^2}{\tilde{\sigma}_i} + \phi \left(\frac{\tilde{\sigma}_i}{\sigma_i} \right) \\ &= \mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m}) + [\mathcal{L}(\hat{s}_m, \hat{\sigma}_m) - \mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m})] \\ &\quad + [\mathcal{L}(\tilde{s}, \tilde{\sigma}) - \mathcal{L}(\hat{s}_m, \hat{\sigma}_m)] + [\mathcal{K}(P_{s,\sigma}, P_{\tilde{s},\tilde{\sigma}}) - \mathcal{L}(\tilde{s}, \tilde{\sigma})]. \end{aligned}$$

By the definition (2.2.3) of \hat{m} , the inequality

$$\mathcal{L}(\tilde{s}, \tilde{\sigma}) - \mathcal{L}(\hat{s}_m, \hat{\sigma}_m) \leq \text{pen}(m) - \text{pen}(\hat{m}) \quad (2.4.7)$$

is true for any $m \in \mathcal{M}$. The difference between the divergence and the likelihood can be expressed as

$$\begin{aligned} \mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m}) - \mathcal{L}(\hat{s}_m, \hat{\sigma}_m) \\ = \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \left(\frac{\sigma_i}{\hat{\sigma}_{m,I}} - 1 \right) \left(1 - \varepsilon_i^{[1]2} \right) \\ - \frac{2(s_i - \hat{s}_{m,i})\sqrt{\sigma_i}\varepsilon_i^{[1]}}{\hat{\sigma}_{m,I}} - \frac{1}{2} \sum_{i=1}^n \left(\varepsilon_i^{[1]2} + \log \sigma_i \right) . \end{aligned} \quad (2.4.8)$$

Using (2.4.7) and (2.4.8), for any $\alpha \in (0, 1)$, we can write

$$\begin{aligned} (1 - \alpha)\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m}) \\ \leq \mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m, \hat{\sigma}_m}) + \text{pen}(m) + G(m) \\ + W_1(\hat{m}) + W_2(\hat{m}) + Z(\hat{m}) - \text{pen}(\hat{m}) \end{aligned} \quad (2.4.9)$$

where, for any $m \in \mathcal{M}$,

$$\begin{aligned} W_1(m) &= \sum_{I \in p_m} \frac{1}{\hat{\sigma}_{m,I}} \left\| \pi_m \Gamma_\sigma^{1/2} \varepsilon^{[1]} \right\|_I^2 , \\ W_2(m) &= \sum_{I \in p_m} \frac{1}{\hat{\sigma}_{m,I}} \left(\left\langle s_m - s, \Gamma_\sigma^{1/2} \varepsilon^{[1]} \right\rangle_I - \frac{\alpha}{2} \|s_m - s\|_I^2 \right) , \\ Z(m) &= \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \left(\left(\frac{\sigma_i}{\hat{\sigma}_{m,I}} - 1 \right) \left(1 - \varepsilon_i^{[1]2} \right) - \alpha \phi \left(\frac{\hat{\sigma}_{m,I}}{\sigma_i} \right) \right) \end{aligned}$$

and

$$G(m) = \sum_{I \in p_m} \left(\frac{1}{\hat{\sigma}_{m,I}} \left\langle s - \hat{s}_m, \Gamma_\sigma^{1/2} \varepsilon^{[1]} \right\rangle_I - \frac{1}{2} \sum_{i \in I} \left(\frac{\sigma_i}{\hat{\sigma}_{m,I}} - 1 \right) \left(1 - \varepsilon_i^{[1]2} \right) \right) .$$

We split the proof of theorem 2.4 in several lemmas.

LEMMA 2.5. *For any $m \in \mathcal{M}$, we have*

$$\mathbb{E}[G(m)] \leq 0 .$$

PROOF. Let us compute this expectation to obtain the inequality. By independence between $\varepsilon^{[1]}$ and $\varepsilon^{[2]}$, we get

$$\begin{aligned} \mathbb{E} \left[G(m) \mid \varepsilon^{[2]} \right] &= \sum_{I \in p_m} -\frac{1}{\hat{\sigma}_{m,I}} \mathbb{E} \left[\left\langle \pi_m \Gamma_\sigma^{1/2} \varepsilon^{[1]}, \Gamma_\sigma^{1/2} \varepsilon^{[1]} \right\rangle_I \right] \\ &= - \sum_{I \in p_m} \frac{1}{\hat{\sigma}_{m,I}} \mathbb{E} \left[\left\| \pi_m \Gamma_\sigma^{1/2} \varepsilon^{[1]} \right\|_I^2 \right] . \end{aligned}$$

It leads to $\mathbb{E}[G(m)] = \mathbb{E}[\mathbb{E}[G(m) \mid \varepsilon^{[2]}]] \leq 0$. □

In order to control $Z(m)$, we split it in two terms that we study separately,

$$Z(m) = Z_+(m) + Z_-(m)$$

where

$$Z_+(m) = \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \left(\left(\frac{\sigma_i}{\hat{\sigma}_{m,I}} - 1 \right)_+ \left(1 - \varepsilon_i^{[1]2} \right) - \alpha \phi \left(\frac{\hat{\sigma}_{m,I}}{\sigma_i} \right) \mathbb{1}_{\hat{\sigma}_{m,I} \leq \sigma_i} \right)$$

and

$$Z_-(m) = \frac{1}{2} \sum_{I \in p_m} \sum_{i \in I} \left(\left(\frac{\sigma_i}{\hat{\sigma}_{m,I}} - 1 \right)_- \left(\varepsilon_i^{[1]2} - 1 \right) - \alpha \phi \left(\frac{\hat{\sigma}_{m,I}}{\sigma_i} \right) \mathbb{1}_{\hat{\sigma}_{m,I} > \sigma_i} \right) .$$

LEMMA 2.6. *Let $m \in \mathcal{M}$ and x be a positive number. Under the hypothesis (\mathbf{H}_θ) , we get*

$$\mathbb{E} [(Z_+(m) - x)_+] \leq \frac{\gamma\theta|p_m|}{\alpha} \exp \left(-\frac{n - (d_m + 3)|p_m|}{2|p_m|} \log \left(1 + \frac{2\alpha|p_m|x}{\gamma n} \right) \right).$$

PROOF. We begin by setting, for all $1 \leq i \leq n$,

$$T_i(m) = \frac{(\sigma_i/\hat{\sigma}_{m,i} - 1)_+}{\left(\sum_{j=1}^n (\sigma_j/\hat{\sigma}_{m,j} - 1)_+^2 \right)^{1/2}}$$

and we denote by

$$S(m) = \sum_{i=1}^n T_i(m) \left(1 - \varepsilon_i^{[1]^2} \right).$$

We lower bound the function ϕ by the remark

$$\forall a \in (0, 1), \forall u \in [a, 1], \left(\frac{1}{u} - 1 \right)^2 \leq \frac{2}{a} \phi(u).$$

Thus, we obtain

$$\sum_{i=1}^n \left(\frac{\sigma_i}{\hat{\sigma}_{m,i}} - 1 \right)_+^2 \leq 2 \left(\max_{i \leq n} \frac{\sigma_i}{\hat{\sigma}_{m,i}} \right) \sum_{j=1}^n \phi \left(\frac{\hat{\sigma}_{m,j}}{\sigma_j} \right) \mathbb{1}_{\hat{\sigma}_{m,j} \leq \sigma_j} = 2M(m)$$

and we use this inequality to get

$$\begin{aligned} Z_+(m) &= \frac{1}{2} \left(\sum_{i=1}^n \left(\frac{\sigma_i}{\hat{\sigma}_{m,i}} - 1 \right)_+^2 \right)^{1/2} S(m) - \frac{\alpha}{2} \sum_{i=1}^n \phi \left(\frac{\hat{\sigma}_{m,i}}{\sigma_i} \right) \mathbb{1}_{\hat{\sigma}_{m,i} \leq \sigma_i} \\ &\leq \sqrt{\frac{M(m)}{2}} S(m)_+ - \frac{\alpha}{2} \sum_{i=1}^n \phi \left(\frac{\hat{\sigma}_{m,i}}{\sigma_i} \right) \mathbb{1}_{\hat{\sigma}_{m,i} \leq \sigma_i} \\ &\leq \frac{1}{4\alpha} \left(\max_{i \leq n} \frac{\sigma_i}{\hat{\sigma}_{m,i}} \right) S(m)_+^2. \end{aligned}$$

To control $S(m)$, we use the inequality (4.2) in [LM00], conditionally to $\varepsilon^{[2]}$. Let $u > 0$,

$$\begin{aligned} \mathbb{P} \left(\left(\max_{i \leq n} \frac{\sigma_i}{\hat{\sigma}_{m,i}} \right) S(m)_+^2 \geq u \right) &= \mathbb{E} \left[\mathbb{P} \left(S(m) \geq \sqrt{u / \max_{i \leq n} \frac{\sigma_i}{\hat{\sigma}_{m,i}}} \mid \varepsilon^{[2]} \right) \right] \\ &\leq \mathbb{E} \left[\exp \left(-\frac{u}{4} \min_{i \leq n} \frac{\hat{\sigma}_{m,i}}{\sigma_i} \right) \right]. \end{aligned}$$

By the remark (2.4.1), we can upper bound it by

$$\mathbb{P} \left(\left(\max_{i \leq n} \frac{\sigma_i}{\hat{\sigma}_{m,i}} \right) S(m)_+^2 \geq u \right) \leq \mathbb{E} \left[\exp \left(-\frac{u}{4\gamma} \min_{I \in p_m} X_I \right) \right]$$

where the X_I 's are *i.i.d.* random variables with a $\chi^2(|I| - d_m - 1)/|I|$ distribution.

For any $\lambda > 0$, we know that the Laplace transform of X_I is given by

$$\mathbb{E} [e^{-\lambda X_I}] = \left(1 + \frac{2\lambda}{|I|} \right)^{-(|I| - d_m - 1)/2}. \quad (2.4.10)$$

Let $t > 0$, the following expectation is dominated by

$$\begin{aligned} \mathbb{E} \left[\left(Z_+(m) - \frac{\gamma}{2\alpha} t \right)_+ \right] &= \int_0^\infty \mathbb{P} \left(Z_+(m) \geq \frac{\gamma t}{2\alpha} + u \right) du \\ &\leq \int_0^\infty \mathbb{E} \left[\exp \left(-\left(\frac{\alpha u}{\gamma} + \frac{t}{2} \right) \min_{I \in p_m} X_I \right) \right] du \\ &\leq \int_0^\infty \mathbb{E} \left[\max_{I \in p_m} \exp \left(-\left(\frac{\alpha u}{\gamma} + \frac{t}{2} \right) X_I \right) \right] du. \end{aligned}$$

Using (\mathbf{H}_θ) and (2.4.10), we roughly upper bound the maximum by the sum of the Laplace transforms and we get

$$\begin{aligned} \mathbb{E} \left[\left(Z_+(m) - \frac{\gamma}{2\alpha} t \right)_+ \right] &\leq \sum_{I \in p_m} \frac{\gamma |I|}{\alpha(|I| - d_m - 3)} \left(1 + \frac{t}{|I|} \right)^{-(|I| - d_m - 3)/2} \\ &\leq \frac{\gamma \theta |p_m|}{\alpha} \exp \left(-\frac{n - (d_m + 3)|p_m|}{2|p_m|} \log \left(1 + \frac{t|p_m|}{n} \right) \right). \end{aligned}$$

Take $t = 2\alpha x / \gamma$ to conclude. \square

LEMMA 2.7. *Let $m \in \mathcal{M}$ and x be a positive number, then*

$$\mathbb{E} [(Z_-(m) - (2\alpha + 1)x)_+] \leq \frac{2\alpha + 1}{\alpha} e^{-\alpha x}.$$

PROOF. Note that for all $u > 1$, we have

$$2\phi(u) \geq \left(\frac{1}{u} - 1 \right)^2.$$

Let $t > 0$, we handle $Z_-(m)$ conditionally to $\varepsilon^{[2]}$ and, using the previous lower bound on ϕ , we obtain

$$\begin{aligned} \mathbb{P} \left(Z_-(m) \geq \frac{2\alpha + 1}{2\alpha} t \mid \varepsilon^{[2]} \right) &\leq \mathbb{P} \left(\frac{1}{2} \sum_{i=1}^n \left(\frac{\sigma_i}{\hat{\sigma}_{m,i}} - 1 \right)_- \left(\varepsilon_i^{[1]2} - 1 \right) \geq \frac{2\alpha + 1}{2\alpha} t + \frac{\alpha}{4} \sum_{i=1}^n \left(\frac{\sigma_i}{\hat{\sigma}_{m,i}} - 1 \right)_-^2 \mid \varepsilon^{[2]} \right) \\ &\leq \mathbb{P} \left(\frac{1}{2} \sum_{i=1}^n \left(\frac{\sigma_i}{\hat{\sigma}_{m,i}} - 1 \right)_- \left(\varepsilon_i^{[1]2} - 1 \right) \geq t + \sqrt{\frac{t}{2} \sum_{i=1}^n \left(\frac{\sigma_i}{\hat{\sigma}_{m,i}} - 1 \right)_-^2} \mid \varepsilon^{[2]} \right). \end{aligned}$$

Let us note that

$$\max_{i \leq n} \left(\frac{\sigma_i}{\hat{\sigma}_{m,i}} - 1 \right)_- \leq 1,$$

thus, we can apply the inequality (4.1) from [LM00] to get

$$\mathbb{P} \left(Z_-(m) \geq \frac{2\alpha + 1}{2\alpha} t \right) \leq \exp(-t/2).$$

This inequality leads us to

$$\begin{aligned} \mathbb{E} \left[\left(Z_-(m) - \frac{2\alpha + 1}{\alpha} t \right)_+ \right] &\leq \int_{(2\alpha+1)t/\alpha}^{+\infty} \mathbb{P}(Z_-(m) \geq u) du \\ &\leq \frac{2\alpha + 1}{\alpha} e^{-t}. \end{aligned}$$

Take $t = \alpha x$ to get the announced result. \square

It remains to control $W_1(m)$ and $W_2(m)$. For the first one, we now prove a Rosenthal-type inequality.

LEMMA 2.8. *Consider any $m \in \mathcal{M}$. Under the hypothesis (\mathbf{H}_θ) , for any $x > 0$, we have*

$$\begin{aligned} \mathbb{E}[(W_1(m) - \gamma \theta D_m - x)_+] &\leq C \theta^2 \gamma \sqrt{|p_m|} d_m \left(\frac{2C \theta^2 \gamma \sqrt{|p_m|} d_m \log(1 + d_m)}{x} \right)^{[2 \log(1 + d_m)]} \end{aligned}$$

where $\lfloor \cdot \rfloor$ is the integral part and C is a positive constant that could be taken equal to

$$C = \frac{12\sqrt{2e}}{\sqrt{e}-1} \approx 43.131 .$$

PROOF. Using the lemma 2.10 and the remark (2.4.1), we dominate $W_1(m)$,

$$W_1(m) \leq W'_1(m) = \gamma \sum_{I \in p_m} \frac{|I|^{d_m}}{|I| - d_m - 1} F_I = \frac{\gamma n d_m}{n - |p_m|(1 + d_m)} \sum_{I \in p_m} F_I$$

where the F_I 's are *i.i.d.* Fisher random variables of parameters $(d_m, n/|p_m| - d_m - 1)$. We denote by F_m the distribution of the F_I 's and we have

$$\frac{\gamma}{2} D_m \leq \gamma |p_m| d_m \leq \mathbb{E}[W'_1(m)] \leq \gamma \theta |p_m| d_m \leq \gamma \theta D_m .$$

Take $x > 0$ and an integer $q > 1$, then

$$\mathbb{E} [(W'_1(m) - \mathbb{E}[W'_1(m)] - x)_+] \leq \frac{\mathbb{E} [(W'_1(m) - \mathbb{E}[W'_1(m)])_+]^q}{(q-1)x^{q-1}} . \quad (2.4.11)$$

We set $V = W'_1(m) - \mathbb{E}[W'_1(m)]$. It is the sum of the independent centered random variables

$$X_I = \frac{\gamma n d_m}{n - |p_m|(1 + d_m)} (F_I - \mathbb{E}[F_I]), \quad I \in p_m .$$

To dominate $\mathbb{E} [V_+^q]$, we use the theorem 9 in [BBLM05]. Let us compute

$$\sum_{I \in p_m} \mathbb{E}[X_I^2] = \frac{2\gamma^2 n^2 d_m (n - 3|p_m|)|p_m|}{(n - |p_m|(d_m + 3))^2 (n - |p_m|(d_m + 5))} \leq 2\gamma^2 \theta^3 |p_m| d_m$$

and so,

$$\mathbb{E} [V_+^q]^{1/q} \leq \sqrt{12\kappa'\gamma^2\theta^3|p_m|d_mq} + q\kappa'\sqrt{2}\mathbb{E} \left[\max_{I \in p_m} |X_I|^q \right]^{1/q}$$

where $\kappa' = \frac{\sqrt{e}}{2(\sqrt{e}-1)}$.

We consider $q = 1 + \lfloor 2 \log(1 + d_m) \rfloor$ where $\lfloor \cdot \rfloor$ is the integral part. For this choice, $q \leq 1 + d_m$ and it implies

$$2|p_m|q < n - |p_m|(1 + d_m) .$$

The hypothesis (\mathbf{H}_θ) allows us to make a such choice. We roughly upper bound the maximum by the sum and we use (\mathbf{H}_θ) to get

$$\begin{aligned} \mathbb{E} \left[\max_{I \in p_m} |X_I|^q \right] &\leq (\gamma \theta d_m)^q \mathbb{E} \left[\max_{I \in p_m} |F_I - \mathbb{E}[F_I]|^q \right] \\ &\leq (\gamma \theta d_m)^q 2^{q-1} (\mathbb{E}[F_m]^q + |p_m| \mathbb{E}[F_m^q]) \\ &\leq \frac{(2\gamma \theta^2 d_m)^q}{2} + \frac{|p_m|}{2} \left(\frac{(2\gamma \theta d_m)(1 + 2(q-1)/d_m)}{1 - 2|p_m|q/(n - |p_m|(1 + d_m))} \right)^q \\ &\leq (6\gamma \theta^2 d_m)^q |p_m| . \end{aligned}$$

Thus, it gives

$$\begin{aligned} \mathbb{E} [V_+^q]^{1/q} &\leq \gamma \theta^2 \left(\sqrt{12\kappa'|p_m|d_mq} + 6\kappa'\sqrt{2}|p_m|^{1/q}d_mq \right) \\ &\leq 6\kappa'\sqrt{2}\gamma \theta^2 \left(\sqrt{|p_m|d_mq} + |p_m|^{1/q}d_mq \right) \\ &\leq 12\kappa'\sqrt{2}\gamma \theta^2 \sqrt{|p_m|d_m} (1 + \lfloor 2 \log(1 + d_m) \rfloor) . \end{aligned}$$

Injecting this inequality in (2.4.11) leads to

$$\begin{aligned} & \mathbb{E} \left[(W_1'(m) - \mathbb{E}[W_1'(m)] - x)_+ \right] \\ & \leq C\gamma\theta^2 \sqrt{|p_m|} d_m \left(\frac{C\gamma\theta^2 \sqrt{|p_m|} d_m (1 + 2\log(1 + d_m))}{2x} \right)^{\lfloor 2\log(1 + d_m) \rfloor}. \end{aligned}$$

□

LEMMA 2.9. Consider any $m \in \mathcal{M}$ and let x be a positive number. Under the hypothesis (\mathbf{H}_θ) , we have

$$\mathbb{E} \left[(W_2(m) - x)_+ \right] \leq \frac{\gamma\theta|p_m|}{\alpha} \exp \left(-\frac{n - (d_m + 3)|p_m|}{2|p_m|} \log \left(1 + \frac{2\alpha|p_m|x}{\gamma n} \right) \right).$$

PROOF. Let us define

$$A(m) = \sum_{I \in p_m} \frac{\|s - s_m\|_I^2}{\hat{\sigma}_{m,I}^2}.$$

The distribution of $W_2(m)$ conditionally to $\varepsilon^{[2]}$ is Gaussian with mean equal to $-\alpha A(m)/2$ and variance factor

$$\sum_{I \in p_m} \frac{\left\| \Gamma_\sigma^{1/2}(s - s_m) \right\|_I^2}{\hat{\sigma}_{m,I}^2}.$$

If ζ is a standard Gaussian random variable, it is well known that, for any $\lambda > 0$,

$$\mathbb{P}(\zeta \geq \sqrt{2\lambda}) \leq e^{-\lambda}. \quad (2.4.12)$$

We apply the Gaussian inequality (2.4.12) to $W_2(m)$ conditionally to $\varepsilon^{[2]}$,

$$\forall t > 0, \mathbb{P} \left(W_2(m) + \frac{\alpha}{2} A(m) \geq \sqrt{2t \sum_{I \in p_m} \frac{\left\| \Gamma_\sigma^{1/2}(s - s_m) \right\|_I^2}{\hat{\sigma}_{m,I}^2}} \middle| \varepsilon^{[2]} \right) \leq e^{-t}.$$

It leads to

$$\mathbb{P} \left(W_2(m) + \frac{\alpha}{2} A(m) \geq \sqrt{2t A(m) \max_{i \leq n} \frac{\sigma_i}{\hat{\sigma}_{m,i}}} \middle| \varepsilon^{[2]} \right) \leq e^{-t}$$

and thus, by the remark (2.4.1),

$$\mathbb{P} \left(W_2(m) \geq \frac{\gamma t}{\alpha} \max_{I \in p_m} X_I^{-1} \middle| \varepsilon^{[2]} \right) \leq \mathbb{P} \left(W_2(m) \geq \frac{t}{\alpha} \max_{i \leq n} \frac{\sigma_i}{\hat{\sigma}_{m,i}} \middle| \varepsilon^{[2]} \right) \leq e^{-t}$$

where the X_I 's are *i.i.d.* random variables with a $\chi^2(|I| - d_m - 1)/|I|$ distribution. Finally, we integrate following $\varepsilon^{[2]}$ and we get

$$\mathbb{P}(W_2(m) \geq t) \leq \mathbb{E} \left[\max_{I \in p_m} \exp \left(-\frac{\alpha t}{\gamma} X_I \right) \right].$$

We finish as we did for $Z_+(m)$,

$$\begin{aligned} & \mathbb{E} \left[\left(W_2(m) - \frac{\gamma}{2\alpha} t \right)_+ \right] \\ & \leq \int_0^{+\infty} \mathbb{E} \left[\max_{I \in p_m} \exp \left(-\left(\frac{\alpha u}{\gamma} + \frac{t}{2} \right) X_I \right) \right] du \\ & \leq \frac{\gamma\theta}{\alpha} \sum_{I \in p_m} \left(1 + \frac{t}{|I|} \right)^{-(|I| - d_m - 3)/2} \\ & \leq \frac{\gamma\theta|p_m|}{\alpha} \exp \left(-\frac{n - (d_m + 3)|p_m|}{2|p_m|} \log \left(1 + \frac{t|p_m|}{n} \right) \right). \end{aligned}$$

□

In order to end the proof of theorem 2.4, we need to put together the results of the previous lemmas. Because $\gamma \geq 1$, for any $x > 0$, we can write

$$e^{-\alpha x} \leq \exp \left(-\frac{n}{2|p_m|} \log \left(1 + \frac{2\alpha|p_m|x}{\gamma n} \right) \right).$$

We now come back to (2.4.9) and we apply the preceding results to each model. Let $m \in \mathcal{M}$, we take

$$x = \frac{x_m}{2(2+\alpha)}$$

and, recalling (2.4.6), we get the following inequalities

$$\begin{aligned} & (1-\alpha)\mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\tilde{s},\tilde{\sigma}})] \\ & \leq \mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\hat{s}_m,\hat{\sigma}_m})] + \text{pen}(m) + \mathbb{E} \left[\left(W_1(\hat{m}) - \gamma\theta D_{\hat{m}} - \frac{x_{\hat{m}}}{2(2+\alpha)} \right)_+ \right] \\ & \quad + \mathbb{E} \left[\left(W_2(\hat{m}) - \frac{x_{\hat{m}}}{2(2+\alpha)} \right)_+ \right] + \mathbb{E} \left[\left(Z_+(\hat{m}) - \frac{x_{\hat{m}}}{2(2+\alpha)} \right)_+ \right] \\ & \quad + \mathbb{E} \left[\left(Z_-(\hat{m}) - (1+2\alpha)\frac{x_{\hat{m}}}{2(2+\alpha)} \right)_+ \right] \\ & \leq \mathbb{E}[\mathcal{K}(m)] + \text{pen}(m) + R_1(\mathcal{M}) + R_2(\mathcal{M}) \end{aligned} \quad (2.4.13)$$

where $R_1(\mathcal{M})$ and $R_2(\mathcal{M})$ are the sums defined in the theorem 2.4. As the choice of m is arbitrary, we can take the infimum among $m \in \mathcal{M}$ in the right part of (2.4.13).

2.4.3. Proof of Proposition 2.3. For the collection \mathcal{F}^{PC} , we have $A = 1$ and $B = 0$ in (2.2.4). Let $m \in \mathcal{M}$, we denote by $\bar{\sigma}_m \in \Sigma_m$ the quantity

$$\bar{\sigma}_m = \sum_{I \in p_m} \bar{\sigma}_{m,I} \mathbb{1}_I \text{ with } \forall I \in p_m, \bar{\sigma}_{m,I} = \frac{1}{|I|} \sum_{i \in I} \sigma_i.$$

The theorem 2.2 gives us

$$\begin{aligned} & \mathbb{E}[\mathcal{K}_n(P_{s,\sigma}, P_{\tilde{s},\tilde{\sigma}})] \\ & \leq \frac{C}{n} \inf_{m \in \mathcal{M}} \{ \mathcal{K}(P_{s,\sigma}, P_{s_m,\sigma_m}) + D_m \log^{1+\epsilon} D_m \} + \frac{R}{n} \\ & \leq \frac{C}{n} \inf_{m \in \mathcal{M}} \{ \mathcal{K}(P_{s,\sigma}, P_{s_m,\bar{\sigma}_m}) + D_m \log^{1+\epsilon} D_m \} + \frac{R}{n} \\ & \leq C \inf_{m \in \mathcal{M}} \left\{ \frac{\|s - s_m\|_2^2}{2n\sigma_*} + \frac{\|\sigma - \bar{\sigma}_m\|_2^2}{2n\sigma_*^2} + D_m \log^{1+\epsilon} D_m \right\} + \frac{R}{n} \end{aligned}$$

because, for any $x > 0$, $\phi(x) \leq (x - 1/x)^2$.

Assuming $(s_r, \sigma_r) \in \mathcal{H}_{\alpha_1}(L_1) \times \mathcal{H}_{\alpha_2}(L_2)$, we know (see [DL93]) that

$$\|s - s_m\|_2^2 \leq nL_1^2(|p_m|d_m)^{-2\alpha_1}$$

and

$$\|\sigma - \bar{\sigma}_m\|_2^2 \leq nL_2^2|p_m|^{-2\alpha_2}.$$

Thus, we obtain

$$\begin{aligned} & \mathbb{E}[\mathcal{K}_n(P_{s,\sigma}, P_{\tilde{s},\tilde{\sigma}})] \\ & \leq C \inf_{m \in \mathcal{M}} \left\{ \frac{L_1^2}{2\sigma_*}(|p_m|d_m)^{-2\alpha_1} + \frac{L_2^2}{2\sigma_*^2}|p_m|^{-2\alpha_2} + \frac{\log^{1+\epsilon} n}{n} D_m \right\} + \frac{R}{n}. \end{aligned}$$

If $\alpha_1 < \alpha_2$, we can take

$$|p_m|d_m = \left\lfloor \left(\frac{L_1^2 n}{2\sigma_* \log^{1+\epsilon} n} \right)^{1/(1+2\alpha_1)} \right\rfloor$$

and

$$|p_m| = \left\lfloor \left(\frac{L_2^2 n}{2\sigma_*^2 \log^{1+\epsilon} n} \right)^{1/(1+2\alpha_2)} \right\rfloor .$$

For $\alpha_1 \geq \alpha_2$, this choice is not allowed because it would imply $d_m = 0$. So, in this case, we take

$$d_m = 1 \text{ and } |p_m| = \left\lfloor \left(\frac{(L_1^2 \sigma_* + L_2^2)n}{2\sigma_*^2 \log^{1+\epsilon} n} \right)^{1/(1+2\alpha_2)} \right\rfloor .$$

In the two situation, we obtain the announced result.

2.5. Technical results

This section is devoted to some useful technical results. Some notations previously introduced can have a different meaning here.

LEMMA 2.10. *Let Σ be a positive symmetric $n \times n$ -matrix and $\sigma_1, \dots, \sigma_n > 0$ be its eigenvalues. Let P be an orthogonal projection of rank $D \geq 1$. If we denote $M = P\Sigma P$, then M is a non-negative symmetric matrix of rank D and, if τ_1, \dots, τ_D are its positive eigenvalues, we have*

$$\min_{1 \leq i \leq n} \sigma_i \leq \min_{1 \leq i \leq D} \tau_i \quad \text{and} \quad \max_{1 \leq i \leq D} \tau_i \leq \max_{1 \leq i \leq n} \sigma_i .$$

PROOF. We denote by $\Sigma^{1/2}$ the symmetric square root of Σ . By a classical result, M has the same rank, equal to D , than $P\Sigma^{1/2}$. On a first side, we have

$$\begin{aligned} \max_{1 \leq i \leq D} \tau_i &= \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\langle P\Sigma P x, x \rangle}{\|x\|^2} \\ &= \sup_{\substack{(x_1, x_2) \in \ker(P) \times \text{Im}(P) \\ (x_1, x_2) \neq (0, 0)}} \frac{\langle P\Sigma x_2, x_2 \rangle}{\|x_1\|^2 + \|x_2\|^2} \\ &\leq \sup_{\substack{x_2 \in \text{Im}(P) \\ x_2 \neq 0}} \frac{\langle \Sigma x_2, x_2 \rangle}{\|x_2\|^2} \leq \max_{1 \leq i \leq n} \sigma_i . \end{aligned}$$

On the other side, we can write

$$\begin{aligned} \min_{1 \leq i \leq D} \tau_i &= \min_{\substack{V \subset \mathbb{R}^n \\ \dim(V)=n-D+1}} \max_{\substack{x \in V \\ x \neq 0}} \frac{\langle Mx, x \rangle}{\|x\|^2} \\ &= \min_{\substack{V \subset \mathbb{R}^n \\ \dim(V)=n-D+1}} \max_{\substack{x \in V \\ x \neq 0}} \frac{\|\Sigma^{1/2} P x\|^2}{\|x\|^2} \\ &\geq \min_{\substack{V \subset \mathbb{R}^n \\ \dim(V)=n-D+1}} \max_{\substack{x \in V \cap \text{Im}(P) \\ x \neq 0}} \frac{\|\Sigma^{1/2} x\|^2}{\|x\|^2} \\ &\geq \min_{\substack{V' \subset \mathbb{R}^n \\ \dim(V') \geq 1}} \max_{\substack{x \in V' \\ x \neq 0}} \frac{\|\Sigma^{1/2} x\|^2}{\|x\|^2} = \min_{1 \leq i \leq n} \sigma_i . \end{aligned}$$

□

LEMMA 2.11. *Let ε be a standard Gaussian vector in \mathbb{R}^n , $a = (a_1, \dots, a_n)' \in \mathbb{R}^n$ and $b_1, \dots, b_n > 0$. We denote by b^* (resp. b_*) the maximum (resp. minimum) of the b_i 's. If $n > 2$ and $Z = \sum_{i=1}^n (a_i + \sqrt{b_i} \varepsilon_i)^2$, then*

$$\mathbb{E} \left[\frac{1}{Z} \right] \leq \frac{1}{\mathbb{E}[Z]} \left(1 + \frac{2\kappa(b^*/b_*)^2}{n-2} \right)$$

where $\kappa > 1$ is a constant that can be taken equal to $1 + 2e^{-1} \approx 1.736$.

PROOF. We recall that $\mathbb{E}[Z] = \sum_{i=0}^n (a_i^2 + b_i)$ and, for any $\lambda > 0$, the Laplace transform of $(a_i + \sqrt{b_i} \varepsilon_i)^2$ is

$$\mathbb{E} \left[\exp \left(-\lambda (a_i + \sqrt{b_i} \varepsilon_i)^2 \right) \right] = \exp \left(-\frac{\lambda a_i^2}{1 + 2\lambda b_i} - \frac{1}{2} \log(1 + 2\lambda b_i) \right).$$

Thus, the Laplace transform of Z is equal to

$$\begin{aligned} \psi(\lambda) &= \mathbb{E} [e^{-\lambda Z}] \\ &= \exp \left(-\sum_{i=1}^n \frac{\lambda a_i^2}{1 + 2\lambda b_i} - \frac{1}{2} \sum_{i=1}^n \log(1 + 2\lambda b_i) \right) \\ &= e^{-\lambda \mathbb{E}[Z]} \times \exp \left(\sum_{i=0}^n \frac{2\lambda^2 a_i^2 b_i}{1 + 2\lambda b_i} - \frac{1}{2} \sum_{i=1}^n r(2\lambda b_i) \right) \end{aligned}$$

where $r(x) = \log(1+x) - x$ for all $x > 0$. To compute the expectation of the inverse of Z , we integrate ψ by parts,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{Z} \right] &= \int_0^\infty \psi(\lambda) d\lambda \\ &= \int_0^\infty e^{-\lambda \mathbb{E}[Z]} \times \exp \left(\sum_{i=0}^n \frac{2\lambda^2 a_i^2 b_i}{1 + 2\lambda b_i} - \frac{1}{2} \sum_{i=1}^n r(2\lambda b_i) \right) d\lambda \\ &= \frac{1}{\mathbb{E}[Z]} + \frac{1}{\mathbb{E}[Z]} \int_0^\infty f_{a,b}(\lambda) \psi(\lambda) d\lambda \end{aligned}$$

where

$$f_{a,b}(\lambda) = \sum_{i=0}^n \frac{2\lambda b_i^2}{1 + 2\lambda b_i} + \frac{4\lambda a_i^2 b_i (1 + \lambda b_i)}{(1 + 2\lambda b_i)^2}.$$

We now upper bound the integral,

$$\begin{aligned} \mathbb{E} \left[\frac{\mathbb{E}[Z]}{Z} - 1 \right] &= \int_0^\infty f_{a,b}(\lambda) \frac{\exp \left(-\sum_{i=1}^n \lambda a_i^2 / (1 + 2\lambda b_i) \right)}{\prod_{i=1}^n \sqrt{1 + 2\lambda b_i}} d\lambda \\ &\leq \int_0^\infty \frac{2n\lambda b_*^2}{(1 + 2\lambda b_*)^{1+n/2}} d\lambda \\ &\quad + \int_0^\infty \frac{4b^*(1 + \lambda b^*)}{(1 + 2\lambda b_*)^{1+n/2}} \times g_{a,b}(\lambda) e^{-g_{a,b}(\lambda)} d\lambda \end{aligned}$$

where we have set

$$g_{a,b}(\lambda) = \sum_{i=1}^n \frac{\lambda a_i^2}{1 + 2\lambda b_i}.$$

For any $t > 0$, $te^{-t} \leq e^{-1}$. Because $g_{a,b}$ is a positive function and $n > 2$, we obtain

$$\begin{aligned}
 \mathbb{E} \left[\frac{\mathbb{E}[Z]}{Z} - 1 \right] &\leq \int_0^\infty \frac{2n\lambda b^{*2}}{(1+2\lambda b_*)^{1+n/2}} d\lambda + \int_0^\infty \frac{4b^*(1+\lambda b^*)}{e(1+2\lambda b_*)^{1+n/2}} d\lambda \\
 &\leq \frac{2(b^*/b_*)^2}{n-2} + \frac{4(b^*/b_*)(n-2+b^*/b_*)}{en(n-2)} \\
 &\leq \frac{2(b^*/b_*)^2}{n-2} \left(1 + \frac{2(n-1)}{en} \right) \\
 &\leq 2(1+2e^{-1}) \frac{(b^*/b_*)^2}{n-2} .
 \end{aligned}$$

□

Estimation of a component in an additive model

RÉSUMÉ. Considérons un vecteur aléatoire $Y \in \mathbb{R}^n$ de moyenne inconnue s et de matrice de covariance $\sigma^2 P_n {}^t P_n$ où P_n est une matrice connue quelconque. L'objet de ce chapitre est l'estimation du vecteur s sous des hypothèses de moment sur les coordonnées de Y ou bien pour un bruit gaussien. Les deux cas sont étudiés pour σ^2 connu ou inconnu. Notre approche ne nécessite aucune hypothèse sur s et est basée sur des méthodes de sélection de modèle non-asymptotiques. Etant donnée une collection $\{S_m, m \in \mathcal{M}\}$ d'espaces linéaires, pour chaque $m \in \mathcal{M}$, nous considérons l'estimateur des moindres carrés \hat{s}_m de s dans S_m . A partir de pénalités non-linéaires en la dimension des modèles, nous proposons un choix de \hat{m} de telle façon que $\hat{s}_{\hat{m}}$ ait un risque quadratique aussi proche que possible du minimum de ceux des estimateurs \hat{s}_m . Des inégalités de type oracle et des résultats d'adaptativité sont prouvés pour $\hat{s}_{\hat{m}}$ d'un point de vue non-asymptotique. Un intérêt particulier est donné au problème d'estimation d'une composante dans un modèle additif. Nous présentons comment nos procédures peuvent être utilisées dans ce cadre. Enfin, les performances de nos estimateurs sont illustrées sur des données simulées.

3.1. Introduction

3.1.1. Additive models. Regression analysis is a very old mathematical subject and the first studies have occurred at the beginning of the 19th century with the works of Legendre [Leg05] and Gauss [Gau09] about estimation of the orbits of astronomical bodies. The general form of a *regression model* can be expressed as

$$Z = f(X) + \sigma\varepsilon \quad (3.1.1)$$

where $X = (X^{(1)}, \dots, X^{(k)})'$ is the k -dimensional vector of *explanatory variables* that belongs to some product space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_k \subset \mathbb{R}^k$, the unknown function $f : \mathcal{X} \rightarrow \mathbb{R}$ is called *regression function*, the positive real number σ is a variance factor and the real random noise ε is such that $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] < \infty$ almost surely. In such a model, we are interested in the behavior of Z in accordance with the fluctuations of X . In other words, we want to explain the random variable Z through the function $f(x) = \mathbb{E}[Z|X = x]$. For this purpose, many approaches have been proposed and, among them, a widely used is the *linear regression*

$$Z = \mu + \sum_{i=1}^k \beta_i X^{(i)} + \sigma\varepsilon \quad (3.1.2)$$

where μ and the β_i are unknown constants. This model benefits from easy interpretation in practice and, from a statistical point of view, allows componentwise analysis. However, a drawback of linear regression is its lack of flexibility for modeling more complex dependencies between Z and the $X^{(i)}$'s. In order to bypass this problem while keeping the advantages of models like (3.1.2), we can generalize them by considering *additive regression models* of the form

$$Z = \mu + \sum_{i=1}^k f_i(X^{(i)}) + \sigma\varepsilon \quad (3.1.3)$$

where the unknown functions $f_i : \mathcal{X}_i \rightarrow \mathbb{R}$ will be referred to as the *components* of the regression function f . The object of this chapter is to construct a data-driven procedure for estimating one of these components on a fixed design (*i.e.* conditionally to some realizations of the random variable X). Our approach is based on nonasymptotic model selection and is free from any prior assumption on f and its components. In particular, we do not make any regularity hypothesis on the function to estimate except to deduce uniform convergence rates for our estimators.

Models (3.1.3) are not new and were first considered in the context of input-output analysis by Leontief [Leo47] and in analysis of variance by Scheffé [Sch59] who called them *additive separable models* and *additive models without interaction* respectively. This kind of model structure is widely used in theoretical economics, in econometric data analysis,... and leads to many well known economic results. For more details about interpretability of additive models in economics, the interested reader could find many references at the end of Chapter 8 of [HMSW04].

As we precise above, regression models are useful for interpreting the effects of X on changes of Z . To this end, the statisticians have to estimate the regression function f . Assuming that we observe a sample $\{(X_1, Z_1), \dots, (X_n, Z_n)\}$ obtained from model (3.1.1), it is well known (see [Sto85]) that the optimal L^2 convergence rate for estimating f is of order $n^{-\alpha/(2\alpha+k)}$ where $\alpha > 0$ is an index of smoothness of f . Note that, for large value of k , this rate becomes slow and the performances of any estimation procedure suffer from what is called the *curse of the dimension* in literature. In this connection, Stone [Sto85] has proved the notable fact that, for additive models (3.1.3), the optimal L^2 rate of convergence for estimating each component f_i of f is the one-dimensional rate $n^{-\alpha/(2\alpha+1)}$. In other terms, estimation of the component f_i in (3.1.3) can be done with the same optimal rate than the one achievable with the model $Z' = f_i(X^{(i)}) + \sigma\varepsilon$. Components estimation in additive models has received a large interest since the eighties and this theory benefited a lot from the the works of Buja *et al.* [BHT89], Hastie and Tibshirani [HT90]. Very popular methods for estimating components in (3.1.3) are based on *backfitting* procedures (see [BF85] for more details). These techniques are iterative and may depend on the starting values. The performances of these methods deeply depends on the choice of some convergence criterion and the nature of the obtained results is usually asymptotic (see, for example, the works of Opsomer and Ruppert [OR97] and Mammen, Linton and Nielsen [MLN99]). More recent noniterative methods have been proposed for estimating marginal effects of the $X^{(i)}$ on the variable Z (*i.e.* how Z fluctuates on average if one explanatory variable is varying while others stay fixed). These procedures, known as *marginal integration estimation*, were introduced by Tjøstheim and Auestad [TA94] and Linton and Nielsen [LN95]. In order to estimate the marginal effect of $X^{(i)}$, these methods take place in two times. First, they estimate the regression function f by a particular estimator f^* , called *pre-smoother*, and then they average f^* according to all the variables except $X^{(i)}$. The way for constructing f^* is fundamental and, in practice, one uses a special kernel estimator (see [RW94] and [SLS99] for a discussion on this subject). To this end, one needs to estimate two unknown bandwidths that are necessary for getting f^* . Dealing with a finite sample, the impact of how we estimate these bandwidths is not clear and, as for backfitting, the theoretical results obtained by these methods are mainly asymptotic.

In contrast with these methods, we are interested here in nonasymptotic procedures to estimate components in additive models. The following subsection of this introduction is devoted to introduce some notations and the framework that we handle but also a short review of existing results in nonasymptotic estimation in additive models.

3.1.2. Statistical framework. We observe a sample $\{(Z_1, (x_1, y_1)), \dots, (Z_n, (x_n, y_n))\}$ where the (x_i, y_i) are distinct deterministic design points with values in some product space $\mathcal{X} \times \mathcal{Y}$. These observations are such that

$$Z_i = s(x_i) + t(y_i) + \sigma\varepsilon_i, \quad i = 1, \dots, n, \quad (3.1.4)$$

where $s : \mathcal{X} \rightarrow \mathbb{R}$ and $t : \mathcal{Y} \rightarrow \mathbb{R}$ are unknown functions, σ is a positive variance factor and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ is an unobservable centered random vector with i.i.d. components of unit variance. This framework is derived from the model (3.1.3) by denoting s for the component that we plan to estimate and t for the sum of μ and the other components. Moreover, to avoid identification problem

in the sequel, we assume

$$\int_{\mathcal{X}} s(x) dx = \int_{\mathcal{Y}} t(y) dy = 0 . \quad (3.1.5)$$

This hypothesis is not restrictive since we are interested in how Z fluctuates with respect to the x_i 's. A shift on the components does not affect these fluctuations and the estimation proceeds up to an additive constant. The results described in this chapter are obtained under two different assumptions on the noise terms ε_i , namely

(H_{Gauß}): the random vector ε is a standard Gaussian vector in \mathbb{R}^n ,

and

(H_{Mom}): the variables ε_i satisfy the moment condition

$$\exists p > 2 \text{ such that } \tau_p = \mathbb{E}[|\varepsilon_1|^p] < \infty . \quad (3.1.6)$$

So, our aim is to estimate the component s on the basis of the observations (3.1.4). For the sake of simplicity of this introduction, we assume that the quantity $\sigma^2 > 0$ is known (see Section 3.5 for unknown variance) and we implicitly identify the functions s and t with the vectors $(s(x_1), \dots, s(x_n))'$ and $(t(y_1), \dots, t(y_n))'$ respectively. Moreover, we assume that we know two linear subspaces $E, F \subset \mathbb{R}^n$ such that $s \in E$, $t \in F$ and $E \oplus F = \mathbb{R}^n$. Of course, such spaces are not available to the statisticians in practice and, when we handle additive models in Section 3.3, we will not suppose that they are known. Let P_n be the projection onto E along F , we derive from (3.1.4) the following regression framework

$$Y = P_n Z = s + \sigma P_n \varepsilon \quad (3.1.7)$$

where $Y = (Y_1, \dots, Y_n)'$ belongs to $E = \text{Im}(P_n) \subset \mathbb{R}^n$. This framework is similar to the classical *signal-plus-noise* regression framework but the data are not independent and their variances are not equal. Because of this uncommonness of the variances of the observations, we qualify (3.1.7) as an *heteroscedastic* framework. The object of this chapter is to estimate the component s and, to this end, we handle (3.1.7). However, the results that we introduce in the sequel consider the framework (3.1.7) from a more general outlook and we do not make any prior hypothesis on P_n . We only assume that it is a projector when we handle the problem of component estimation in an additive framework.

We now describe our estimation procedure in details. For any $z \in \mathbb{R}^n$, we define the *least-squares contrast* by

$$\gamma_n(z) = \|Y - z\|_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - z_i)^2 . \quad (3.1.8)$$

Let us consider a collection of linear subspaces of $\text{Im}(P_n)$ denoted by $\mathcal{F} = \{S_m, m \in \mathcal{M}\}$ where \mathcal{M} is a finite or countable index set. Hereafter, the S_m 's will be called the *models*. Denoting by π_m the orthogonal projection onto S_m , the minimum of γ_n over S_m is achieved at a single point $\hat{s}_m = \pi_m Y$ called the *least-squares estimator* of s in S_m . Note that the expectation of \hat{s}_m is equal to the orthogonal projection $s_m = \pi_m s$ of s onto S_m . We have the following identity for the quadratic risk of \hat{s}_m 's,

PROPOSITION 3.1. *Let $m \in \mathcal{M}$, the least-squares estimator $\hat{s}_m = \pi_m Y$ of s on S_m satisfies*

$$\mathbb{E}[\|s - \hat{s}_m\|_n^2] = \|s - s_m\|_n^2 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 . \quad (3.1.9)$$

PROOF. By orthogonality, we have

$$\|s - \hat{s}_m\|_n^2 = \|s - s_m\|_n^2 + \sigma^2 \|\pi_m P_n \varepsilon\|_n^2 . \quad (3.1.10)$$

Because the components of ε are independent and centered with unit variance, we easily compute

$$\mathbb{E}[\|\pi_m P_n \varepsilon\|_n^2] = \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} .$$

We conclude by taking the expectation on both side of (3.1.10). \square

A “good” estimator is such that its quadratic risk is small. The decomposition given by (3.1.9) shows that this risk is a sum of two terms that can be interpreted as follows. The first one, called *bias term*, corresponds to the capacity of the model S_m to approximate the true value of s . The second, called *variance term*, is proportional to $\text{Tr}(^tP_n\pi_mP_n)$ and measures, in a certain sense, the complexity of S_m . Indeed, if P_n is the identity matrix, for example, then $\text{Tr}(^tP_n\pi_mP_n)$ is equal to the dimension of S_m . If S_m is the space of constant vectors of \mathbb{R}^n , then the variance term is small but the bias term is as large as s is far from a constant. Conversely, if S_m is a “huge” model, whole \mathbb{R}^n for instance, the bias is null but the price is a great variance term. Thus, the formula (3.1.9) illustrate why choosing a “good” model corresponds to find a trade-off between bias and variance terms.

In the particular case of Gaussian errors with possibly unknown variance factor σ^2 and invertible matrix P_n , another popular way to proceed is to minimize some likelihood criterion. Such approach leads to the maximum likelihood estimator given by

$$\hat{s}_m^L = P_n\pi_m^{(P_n)}P_n^{-1}Y$$

where $\pi_m^{(P_n)}$ is the orthogonal projection onto $P_n^{-1}S_m$. Thanks to the invertibility of P_n , the basic idea to estimate s in this case is to consider the classical homoscedastic data given by $P_n^{-1}Y$. The mean vector $P_n^{-1}s$ is estimated in the models of the collection $\{P_n^{-1}S_m, m \in \mathcal{M}\}$ and the estimators are next transposed to the S_m 's. For simultaneous estimation of s and σ^2 with the Kullback risk in such a framework, we refer to the works of Baraud, Giraud and Huet [BGH09]. Indeed, their procedures can easily be generalized to any invertible matrix by the slight change on the collection of models precised above. The quadratic risk of the maximum likelihood estimator can be computed likewise the one of \hat{s}_m and we obtain a similar decomposition

$$\mathbb{E}[\|s - \hat{s}_m^L\|_n^2] = \|s - P_n\pi_m^{(P_n)}P_n^{-1}s\|_n^2 + \frac{\text{Tr}(P_n\pi_m^{(P_n)}{}^tP_n)}{n}\sigma^2. \quad (3.1.11)$$

Note that the orders of the variance terms in the risk of \hat{s}_m and \hat{s}_m^L are similar. Since $P_n\pi_m^{(P_n)}P_n^{-1}s \in S_m$, the bias term of \hat{s}_m is always no larger than the one of \hat{s}_m^L . Our present aim is to construct estimators with a small quadratic risk and thus, in the sequel of this chapter, we only focus on the least-squares estimators \hat{s}_m . For a further discussion, see Appendix A that is devoted to illustrate the behaviours of the quadratic risks of \hat{s}_m and \hat{s}_m^L on an example.

Clearly, the choice of a model that minimizes the risk (3.1.9) depends on the unknown vector s and make good models unavailable to the statisticians. So, we need a data-driven procedure to select an index $\hat{m} \in \mathcal{M}$ such that $\mathbb{E}[\|s - \hat{s}_{\hat{m}}\|_n^2]$ is close to the smaller L^2 risk among the collection of estimators $\{\hat{s}_m, m \in \mathcal{M}\}$, namely

$$\mathcal{R}(s, \mathcal{F}) = \inf_{m \in \mathcal{M}} \mathbb{E}[\|s - \hat{s}_m\|_n^2].$$

To choose such a \hat{m} , a classical way in model selection consists in minimizing an empirical penalized criterion stochastically close to the risk. Given a *penalty* function $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+$, we define \hat{m} as any minimizer over \mathcal{M} of the penalized least-squares criterion

$$\hat{m} \in \underset{m \in \mathcal{M}}{\text{argmin}} \{\gamma_n(\hat{s}_m) + \text{pen}(m)\}. \quad (3.1.12)$$

This way, we select a model $S_{\hat{m}}$ and we have at our disposal the *penalized least-squares estimator* $\tilde{s} = \hat{s}_{\hat{m}}$. Note that, by definition, the estimator \tilde{s} satisfies

$$\forall m \in \mathcal{M}, \gamma_n(\tilde{s}) + \text{pen}(\hat{m}) \leq \gamma_n(\hat{s}_m) + \text{pen}(m). \quad (3.1.13)$$

To study the performances of \tilde{s} , we have in mind to upperbound its quadratic risk. To this end, we establish inequalities of the form

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq C \inf_{m \in \mathcal{M}} \{\|s - s_m\|_n^2 + \text{pen}(m)\} + \frac{R}{n} \quad (3.1.14)$$

where C and R are numerical terms that do not depend on n . Note that if the penalty is proportional to $\text{Tr}(^tP_n\pi_mP_n)\sigma^2/n$, then the quantity involved in the infimum is of order of the L^2 risk of \hat{s}_m .

Consequently, under suitable assumptions, such inequalities allows us to deduce upperbounds of order of the minimal risk among the collection of estimators $\{\hat{s}_m, m \in \mathcal{M}\}$. This result is known as an *oracle inequality*

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] \leq C\mathcal{R}(s, \mathcal{F}) = C \inf_{m \in \mathcal{M}} \mathbb{E} [\|s - \hat{s}_m\|_n^2] . \quad (3.1.15)$$

This kind of procedure is not new and the first results in estimation by penalized criterion are due to Akaike [Aka70] and Mallows [Mal73] in the early seventies. Since these works, model selection has known an important development and it would be beyond the scope of this chapter to make an exhaustive historical review of the domain. We refer to the first chapters of [MT98] for a more general introduction. Nonasymptotic model selection approach for estimating components in an additive model was studied in few paper only. Considering penalties that are linear in the dimension of the models, Baraud, Comte and Viennet [BCV01] have obtained general results for geometrically β -mixing regression models. Applying it to the particular case of additive models, they estimate the regression function. They obtain nonasymptotic upperbound similar to (3.1.14) on condition ε admits a moment of order larger than 6. For additive regression on a random design and alike penalties, Baraud [Bar02] proved oracle inequalities on the estimation of the regression function for polynomial collection of models and a noise that admits a moment of order 4. Recently, Brunel and Comte [BC06] have obtained results with the same flavor for the estimation of the regression function in an censored additive model and a noise admitting a moment of order larger than 8. Pursuant to this work, Brunel and Comte [BC08] have also proposed a nonasymptotic iterative method to achieve the same goal. Combining ideas from sparse linear modeling and additive regression, Ravikumar *et al.* [RLLW08] have recently developed a data-driven procedure, called SpAM, for estimating a sparse high-dimensional regression function. Some of their empirical results have been proved by Meier, van de Geer and Bühlmann [MvB09] in the case of a sub-Gaussian noise and some sparsity-smoothness penalty.

The methods that we use are similar to the ones of Baraud *et al.*, Comte and Viennet and are inspired from [Bar00]. Nevertheless, the objects of the procedure differ. The works cited above are all connected to the estimation of the whole regression function by estimating simultaneously all its components. Since these components are each treated in the same way, their procedures can not focus on the properties of one of them. The procedure that we purpose in this chapter estimates one of the components and is based on penalties that are not linear in the dimension of the models. Moreover, under mild assumptions on \mathcal{F} , we obtain oracle inequalities under Gaussian assumption on the noise or only under a weak moment condition. Then, we deduce uniform convergence rates over Hölderian balls and adaptivity of our estimators. Up to the best of our knowledge, our results in nonasymptotic estimation of a nonparametric component in an additive regression model are new.

The chapter is organized as follows. In the section 3.2, we study the properties of the estimation procedure under the hypotheses $(\mathbf{H}_{\text{Gauß}})$ and $(\mathbf{H}_{\text{Mom}})$ with a known variance factor σ^2 . As a consequence, we deduce oracle inequalities and we discuss about the size of the collection of models. In Section 3.3, we apply these results to the particular case of the additive models and, in the next section, we give rates of convergence for our estimators over Hölderian balls. The case of unknown variance factor is presented in Section 3.5 and the results of the first section are extended to this situation. Finally, in Section 3.6, we illustrate the performances of our estimators in practice by a simulation study. The last sections are devoted to the proofs and to some technical lemmas.

Notations: in the sequel, we denote by ρ the *spectral norm* on \mathbb{M}_n as the norm induced by $\|\cdot\|_n$,

$$\forall A \in \mathbb{M}_n, \rho(A) = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_n}{\|x\|_n} . \quad (3.1.16)$$

In particular, for any $A, B \in \mathbb{M}_n$, the norm ρ satisfies these properties:

1. $\rho(AB) \leq \rho(A)\rho(B)$ (sub-multiplicativity),
2. $\rho(A) = \rho({}^tA)$ (self-adjoint),
3. $\rho({}^tAA) = \rho(A{}^tA) = \rho^2(A)$,

4. $\rho(A) = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } {}^tAA\}$.

For more details on ρ , see Chapter 5 of [HJ90].

3.2. Main results

Throughout this section, we assume that the variance factor σ^2 in (3.1.7) is known. Moreover, in the sequel of this chapter, for any $d \in \mathbb{N}$, we define N_d as the number of models of dimension d in \mathcal{F} ,

$$N_d = \text{Card}\{m \in \mathcal{M} : \dim(S_m) = d\}.$$

We first introduce general model selection theorems under hypotheses $(\mathbf{H}_{\text{Gau\ss}})$ and $(\mathbf{H}_{\text{Mom}})$.

THEOREM 3.2. *Assume that $(\mathbf{H}_{\text{Gau\ss}})$ holds and consider a collection of nonnegative numbers $\{L_m, m \in \mathcal{M}\}$. Let $\theta > 0$, if the penalty function is such that*

$$\text{pen}(m) \geq (1 + \theta + L_m) \frac{\text{Tr}({}^tP_n \pi_m P_n)}{n} \sigma^2 \text{ for all } m \in \mathcal{M}, \quad (3.2.1)$$

then the penalized least-squares estimator \tilde{s} given by (3.1.12) satisfies

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] \leq \left(1 + \frac{1}{\theta}\right) \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \text{pen}(m) - \frac{\text{Tr}({}^tP_n \pi_m P_n)}{n} \sigma^2 \right\} + \frac{\rho^2(P_n) \sigma^2}{n} R_n(\theta) \quad (3.2.2)$$

where we have set

$$R_n(\theta) = \frac{2(1+\theta)^4}{\theta^3} \sum_{m \in \mathcal{M}} \exp \left(-\frac{\theta^2 L_m}{2(1+\theta)^3} \times \frac{\text{Tr}({}^tP_n \pi_m P_n)}{\rho^2(P_n)} \right).$$

If the errors are not supposed to be Gaussian but only to satisfy the moment condition $(\mathbf{H}_{\text{Mom}})$, the following upperbound on the q th moment of $\|s - \tilde{s}\|_n^2$ holds.

THEOREM 3.3. *Assume that $(\mathbf{H}_{\text{Mom}})$ holds and take $q > 0$ such that $2(q+1) < p$. Consider $\theta > 0$ and some collection $\{L_m, m \in \mathcal{M}\}$ of positive weights. If the penalty function is such that*

$$\text{pen}(m) \geq (1 + \theta + L_m) \frac{\text{Tr}({}^tP_n \pi_m P_n)}{n} \sigma^2 \text{ for all } m \in \mathcal{M}, \quad (3.2.3)$$

then the penalized least-squares estimator \tilde{s} given by (3.1.12) satisfies

$$\mathbb{E} [\|s - \tilde{s}\|_n^{2q}]^{1/q} \leq C \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \text{pen}(m) \right\} + \frac{\rho^2(P_n) \sigma^2}{n} R_n(p, q, \theta)^{1/q} \quad (3.2.4)$$

where we have set

$$R_n(p, q, \theta) = C' \tau_p \left[N_0 + \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \left(1 + \frac{\text{Tr}({}^tP_n \pi_m P_n)}{\rho^2(\pi_m P_n)} \right) \left(\frac{L_m \text{Tr}({}^tP_n \pi_m P_n)}{\rho^2(P_n)} \right)^{q-p/2} \right]$$

and $C = C(q, \theta)$, $C' = C'(p, q, \theta)$ are positive constants.

The proofs of these theorems give explicit values for the constants C that appear in the upperbounds. In both cases, this constant goes to infinity as θ tends to 0 or increases toward infinity. In practice, it does neither seem reasonable to choose θ close to 0 nor very large. Thus this explosive behavior is not restrictive but we still have to choose a “good” θ . The values for θ suggested by the proofs are around the unity but we make no claim of optimality. Indeed, this is a hard problem to determine an optimal choice for θ from theoretical computations since it could depend on all the parameters and on the choice of the collection of models. A solution to calibrate it in a particular case could be a simulation study.

For penalties of order of $\text{Tr}({}^tP_n \pi_m P_n) \sigma^2 / n$, Inequalities (3.2.2) and (3.2.4) are not far from being oracle. Let us denote by R_n the remainder term $R_n(\theta)$ or $R_n(p, q, \theta)$ according to $(\mathbf{H}_{\text{Gau\ss}})$ or $(\mathbf{H}_{\text{Mom}})$ holds. To deduce oracle inequalities from that, we need some additional hypotheses as the following ones:

(**A**₁): there exists some universal constant $K > 0$ such that

$$\text{pen}(m) \leq K \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2, \text{ for all } m \in \mathcal{M},$$

(**A**₂): there exists some constant $R > 0$ such that

$$\sup_{n \geq 1} R_n \leq R,$$

(**A**₃): there exists some constant $\rho > 1$ such that

$$\sup_{n \geq 1} \rho^2(P_n) \leq \rho^2.$$

Thus, under the hypotheses of Theorem 3.2 and these three assumptions, we deduce from (3.2.2) that

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] \leq C \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \right\} + \frac{R \rho^2 \sigma^2}{n} \quad (3.2.5)$$

where C is a constant that does not depend on s , σ^2 and n . By Proposition 3.1, this inequality corresponds to (3.1.15) up to some additive term. To derive similar inequality from (3.2.4), we need on top of that to assume that $p > 4$ in order to be able to take $q = 1$.

Assumption (**A**₃) is subtle and strongly depends on the nature of P_n . The case of oblique projector that we use to estimate a component in an additive framework will be discussed in section 3.3. Let us replace it, for the moment, by the following one

(**A**₃[']): there exists some factor $c \in (0, 1)$ that does not depend on n such that

$$c \rho^2(P_n) \dim(S_m) \leq \text{Tr}(^t P_n \pi_m P_n).$$

By the properties of the norm ρ , note that $\text{Tr}(^t P_n \pi_m P_n)$ admits an upperbound with the same flavor

$$\begin{aligned} \text{Tr}(^t P_n \pi_m P_n) &= \text{Tr}(\pi_m P_n {}^t(\pi_m P_n)) \\ &\leq \rho(\pi_m P_n {}^t(\pi_m P_n)) \text{rk}(\pi_m P_n {}^t(\pi_m P_n)) \\ &\leq \rho^2(\pi_m P_n) \text{rk}(\pi_m) \\ &\leq \rho^2(P_n) \dim(S_m). \end{aligned}$$

In all our results, the quantity $\text{Tr}(^t P_n \pi_m P_n)$ stands for a dimensional term relative to S_m . Hypothesis (**A**₃[']) formalizes that by assuming that its order is the dimension of the model S_m up to the norm of the covariance matrix $^t P_n P_n$.

Let us now discuss about the assumptions (**A**₁) and (**A**₂). They are connected and they raise the impact of the complexity of the collection \mathcal{F} on the estimation procedure. Typically, condition (**A**₂) will be fulfilled under (**A**₁) when \mathcal{F} is not too “large”, that is, when the collection does not contain too many models with the same dimension. We illustrate this phenomenon by the two following corollaries.

COROLLARY 3.4. *Assume that (**H**_{Gauß}) and (**A**₃[']) hold and consider some finite $A \geq 0$ such that*

$$\sup_{d \in \mathbb{N}: N_d > 0} \frac{\log N_d}{d} \leq A. \quad (3.2.6)$$

Let L , θ and ω be some positive numbers that satisfy

$$L \geq \frac{2(1+\theta)^3}{c\theta^2} (A + \omega). \quad (3.2.7)$$

Then, the estimator \tilde{s} obtained from (3.1.12) with penalty function given by

$$\text{pen}(m) = (1 + \theta + L) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2$$

is such that

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] \leq C \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + (L \vee 1) \frac{\text{Tr}(^t P_n \pi_m P_n) \vee (c\rho^2(P_n))}{n} \sigma^2 \right\}$$

where $C > 1$ only depends on θ , ω and c .

For errors that only satisfy moment condition, we have the following similar result.

COROLLARY 3.5. Assume that $(\mathbf{H}_{\text{Mom}})$ and (\mathbf{A}'_3) hold with $p > 6$ and let $A > 0$ and $\omega > 0$ such that

$$N_0 \leq 1 \quad \text{and} \quad \sup_{d > 0: N_d > 0} \frac{N_d}{(1 + d)^{p/2 - 3 - \omega}} \leq A. \quad (3.2.8)$$

Consider some positive numbers L , θ and ω' that satisfy

$$L \geq \omega' A^{2/(p-2)},$$

then, the estimator \tilde{s} obtained from (3.1.12) with penalty function given by

$$\text{pen}(m) = (1 + \theta + L) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2$$

is such that

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] \leq C \tau_p \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + (L \vee 1) \frac{\text{Tr}(^t P_n \pi_m P_n) \vee (c\rho^2(P_n))}{n} \sigma^2 \right\}$$

where $C > 1$ only depends on θ , p , ω , ω' and c .

Note that the assumption (\mathbf{A}'_3) guarantees that $\text{Tr}(^t P_n \pi_m P_n) \geq c\rho^2(P_n) \dim(S_m)$ and, at least for the models with positive dimension, this implies $\text{Tr}(^t P_n \pi_m P_n) \geq c\rho^2(P_n)$. Consequently, up to the factor L , the upperbounds of $\mathbb{E} [\|s - \tilde{s}\|_n^2]$ given by Corollaries 3.4 and 3.5 are of order of the minimal risk $\mathcal{R}(s, \mathcal{F})$. To deduce oracle inequalities for \tilde{s} from that, (\mathbf{A}_1) needs to be fulfilled. In other terms, we need to be able to consider some L independently from the size n of the data. It will be the case if the same is true for the bounds A .

Let us assume that the collection \mathcal{F} is small in the sense that, for any $d \in \mathbb{N}$, the number of models N_d is bounded by some constant term that neither depends on n nor d . Typically, collections of nested models satisfy that. In this case, we are free to take L equal to some universal constant. So, (\mathbf{A}_1) is true for $K = 1 + \theta + L$ and oracle inequalities can be deduced for \tilde{s} . Conversely, a large collection \mathcal{F} is such that there are many models with the same dimension. We consider that this situation happens, for example, when the order of A is $\log n$. In such a case, we need to choose L of order $\log n$ too and the upperbounds on the risk of \tilde{s} become oracle type inequalities up to some logarithmic factor. However, we know that in some situation, this factor can not be avoided as in the complete variable selection problem with Gaussian errors (see Chapter 4 of [Mas07]).

More generally, note that the assumption on N_d in Corollary 3.5 is more restrictive than the one in Corollary 3.4. Indeed, in the Gaussian case, the quantity N_d is limited by e^{Ad} while the bound is only polynomial in d under moment condition. Thus, the Gaussian assumption $(\mathbf{H}_{\text{Gauß}})$ allows to obtain oracle inequalities for more general collections of models.

3.3. Application to additive models

In this section, we focus on the framework (3.1.4) given by an additive model. To describe the procedure to estimate the component s , we assume that the variance factor σ^2 is known but it can be easily generalized to the unknown factor case by considering the results of the section 3.5. Let $\{(x_1, y_1), \dots, (x_n, y_n)\}$ be deterministic design points of $[-1, 1]^2$ and s and t be two unknown functions that belong to $L^2([-1, 1], dx)$ and that satisfy (3.1.5). We recall that we observe

$$Z_i = s(x_i) + t(y_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

where the random vector $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ is such that $(\mathbf{H}_{\text{GauB}})$ or $(\mathbf{H}_{\text{Mom}})$ holds. For legibility, we identify the functions s and t with the vectors $(s_1, \dots, s_n)'$ and $(t_1, \dots, t_n)'$, respectively, where $s_i = s(x_i)$ and $t_i = t(y_i)$.

Let \mathcal{S}_n and \mathcal{S}'_n be two linear subspaces of $L^2([-1, 1], dx)$ with finite dimension $D_n = \dim(\mathcal{S}_n)$ and $D'_n = \dim(\mathcal{S}'_n)$ such that $D_n + D'_n \leq n$. We consider two orthonormal bases $\{\phi_1, \dots, \phi_{D_n}\}$ and $\{\psi_1, \dots, \psi_{D'_n}\}$ of \mathcal{S}_n and \mathcal{S}'_n respectively. The linear spans $E, F \subset \mathbb{R}^n$ are defined by

$$E = \text{Span} \{(\phi_i(x_1), \dots, \phi_i(x_n))', i = 1, \dots, D_n\}$$

and

$$F = \text{Span} \{(\psi_i(y_1), \dots, \psi_i(y_n))', i = 1, \dots, D'_n\}.$$

We make the mild assumption that $E \cap F = \{0\}$. Note that we do not assume that s belongs to E neither that t belongs to F . Let G be the space $(E + F)^\perp$, we obviously have $E \oplus F \oplus G = \mathbb{R}^n$ and we denote by P_n the projection onto E along $F + G$. Moreover, we define π_E and π_{F+G} the orthogonal projections onto E and $F + G$ respectively. Thus, we derive the following framework from (3.1.4),

$$Y = P_n Z = \bar{s} + \sigma P_n \varepsilon \quad (3.3.1)$$

where we have set

$$\begin{aligned} \bar{s} &= P_n s + P_n t \\ &= s + (P_n - I_n)s + P_n t \\ &= s + (P_n - I_n)(s - \pi_E s) + P_n(t - \pi_{F+G} t) = s + h. \end{aligned}$$

Let $\mathcal{F} = \{S_m, m \in \mathcal{M}\}$ be a finite collection of linear subspaces of E , we apply the procedure described in the previous sections to Y given by (3.3.1), that is, we choose an index $\hat{m} \in \mathcal{M}$ as a minimizer of (3.1.12) with a penalty function satisfying the hypotheses of Theorems 3.2 or 3.3 according to $(\mathbf{H}_{\text{GauB}})$ or $(\mathbf{H}_{\text{Mom}})$ holds. This way, we estimate s by \tilde{s} . From the triangular inequality we know

$$\|s - \tilde{s}\|_n \leq \|\bar{s} - \tilde{s}\|_n + \|h\|_n$$

and we derive that

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq 2\mathbb{E}[\|\bar{s} - \tilde{s}\|_n^2] + 2\|h\|_n^2.$$

As we discussed previously, under suitable assumptions on the complexity of the collection \mathcal{F} , we can assume that (\mathbf{A}_1) and (\mathbf{A}_2) are fulfilled. Moreover, we suppose in this section that (\mathbf{A}_3) is satisfied. Note that, for any $m \in \mathcal{M}$, π_m is an orthogonal projection onto the image set of the oblique projection P_n . Consequently, we have $\text{Tr}(^t P_n \pi_m P_n) \geq \text{rk}(\pi_m) = \dim(S_m)$ and Assumption (\mathbf{A}_3) implies (\mathbf{A}'_3) with $c = 1/\rho^2$. Since, for all $m \in \mathcal{M}$,

$$\|\bar{s} - \pi_m \bar{s}\|_n \leq \|s - \pi_m s\|_n + \|h - \pi_m h\|_n \leq \|s - \pi_m s\|_n + \|h\|_n,$$

we deduce from Theorems 3.2 or 3.3 that we can find, independently from s and n , two positive numbers C and C' such that

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq C \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \right\} + C' \left(\|h\|_n^2 + \frac{\rho^2 \sigma^2}{n} R \right). \quad (3.3.2)$$

To derive an interesting upperbound on the L^2 risk of \tilde{s} , we need to control the remainder term. Because $\rho(\cdot)$ is a norm on \mathbb{M}_n , we dominate the norm of h by

$$\begin{aligned} \|h\|_n &\leq \rho(I_n - P_n) \|s - \pi_E s\|_n + \rho(P_n) \|t - \pi_{F+G} t\|_n \\ &\leq (1 + \rho(P_n)) (\|s - \pi_E s\|_n + \|t - \pi_{F+G} t\|_n) \\ &\leq (1 + \rho) (\|s - \pi_E s\|_n + \|t - \pi_{F+G} t\|_n). \end{aligned}$$

Note that, for any $m \in \mathcal{M}$, $S_m \subset E$ and so, $\|s - \pi_E s\|_n \leq \|s - \pi_m s\|_n$. Thus, Inequality (3.3.2) leads to

$$\begin{aligned} \mathbb{E}[\|s - \tilde{s}\|_n^2] &\leq C(1 + \rho)^2 \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \right\} \\ &\quad + C'(1 + \rho)^2 \left(\|t - \pi_{F+G} t\|_n^2 + \frac{\sigma^2}{n} R \right). \end{aligned} \quad (3.3.3)$$

Under assumption on the regularity of the component t , the quantity $\|t - \pi_{F+G}t\|_n^2$ is reasonably small. It mainly remains to understand the order of the multiplicative factor $(1 + \rho)^2$ in practice.

Thus, we now discuss about the norm $\rho(P_n)$ and the assumption **(A₃)**. This quantity depends on how we construct the spaces E and F , *i.e.* on the choice of the functions ϕ_i and ψ_i and the structures of \mathcal{S}_n and \mathcal{S}'_n inherited from these bases. Let us consider the two following particular sets of functions.

(T): Assume that $D = D_n = D'_n$ are even integers. The spaces \mathcal{S}_n and \mathcal{S}'_n are generated by the trigonometric basis on $[-1, 1]$ given by the functions

$$\phi_{2k}(t) = \psi_{2k}(t) = \sin(k\pi t), \quad k = 1, \dots, D/2$$

and

$$\phi_{2k-1}(t) = \psi_{2k-1}(t) = \cos(k\pi t), \quad k = 1, \dots, D/2.$$

(P_r): Let K be a positive integer and consider the regular partition $\{I_1, \dots, I_K\}$ of $[-1, 1]$. For some positive integer r , the functions ϕ_i and ψ_i are polynomials of degree less than r on one of the I_j and zero outside such that \mathcal{S}_n and \mathcal{S}'_n are spaces of piecewise polynomials functions of the form

$$f(t) = \sum_{j=1}^K P_j(t) \mathbb{1}_{I_j}, \quad t \in [-1, 1],$$

where the P_j 's are polynomials of degree less than r given by linear combinations of the ϕ_i (resp. ψ_i) for \mathcal{S}_n (resp. \mathcal{S}'_n). In this case, we have $D_n, D'_n \leq (r+1)K$.

As we described above, the spaces \mathcal{S}_n and \mathcal{S}'_n are constructed on the design given by the points $(x_1, y_1), \dots, (x_n, y_n) \in [-1, 1]^2$. Hereafter, the (x_i, y_i) will be assumed to be known realizations of an uniform random variable on the square $[-1, 1]^2$. In other terms, these points are random and we proceed conditionally to them. We have to choose D_n and D'_n with a good order for ensuring that **(A₃)** occurs for some reasonable value ρ (*i.e.* of order of one). Let us illustrate it by the figures 3.1 and 3.2 that show estimations of the fluctuations of the mean value of $\rho(P_n)$ according to n for several choices of D_n (D'_n is of the same order) in the frameworks obtained from **(T)** and **(P₂)**. For getting a $\rho(P_n)$ close to some constant, these simulations suggest empirical upperbounds on D_n that are similar to the ones given by Baraud [Bar02]. Indeed, $\rho(P_n)$ seems to be harder to control for a basis as **(T)** than for a localized basis like **(P₂)**. In the trigonometric case, taking D_n of order of $\sqrt{n/\log(n)}$ appears to be a reasonable choice to allows ρ to be close to 1. On Figure 3.2, we see that the polynomial case is more flexible and let us take D_n of order \sqrt{n} to empirically satisfy condition **(A₃)**. These bounds are not theoretical but give good results in practice as it is illustrated in Section 3.6.

3.4. Convergence rates

The previous sections have introduced various upperbounds on the L^2 risk of the penalized least-squares estimators \tilde{s} . Each of them is connected to the minimal risk of the estimators among a collection $\{\hat{s}_m, m \in \mathcal{M}\}$. One of the main advantages of such inequalities is that it allows us to derive uniform convergence rates with respect to many well known classes of smoothness (see [BM97]). In this section, we give such results over Hölderian balls for the estimation of a component in an additive framework. To this end, for any $\alpha \in (0, 1)$ and $L > 0$, we introduce the space $\mathcal{H}_\alpha(L)$ of the α -Hölderian functions with constant L on $[-1, 1]$,

$$\mathcal{H}_\alpha(L) = \{f : [-1, 1] \rightarrow \mathbb{R} : \forall x, y \in [-1, 1], |f(x) - f(y)| \leq L|x - y|^\alpha\}.$$

Considering the framework (3.1.4), we define the projector P_n constructed *via* the basis **(P₂)** with $D_n = 3 \times 2^{k_n}$ for some positive integer k_n . By applying P_n to the data vector $Z = (Z_1, \dots, Z_n)'$,

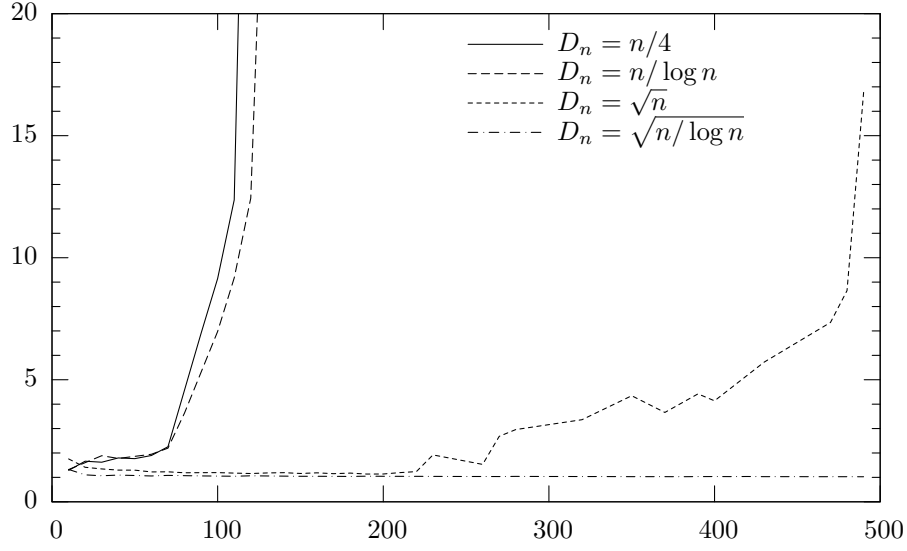


FIGURE 3.1. Estimation of $\rho(P_n)$ for random design points uniformly taken in $[-1, 1]^2$ and collection of functions (\mathbf{T})

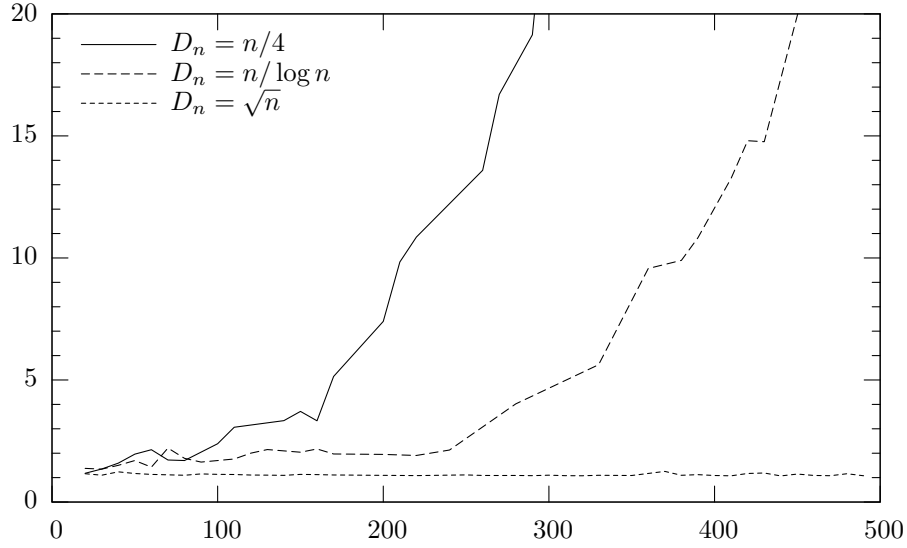


FIGURE 3.2. Estimation of $\rho(P_n)$ for random design points uniformly taken in $[-1, 1]^2$ and collection of functions (\mathbf{P}_2)

we handle the framework (3.3.1) and we have in mind to estimate the component s . The image set of P_n is the space of piecewise polynomials of degree 2 constructed on the regular partition of $[-1, 1]$ whose blocks are given by

$$I_i = \left[-1 + \frac{2(i-1)}{2^{k_n}}, -1 + \frac{2i}{2^{k_n}} \right], \quad i = 1, \dots, 2^{k_n}.$$

Let \mathcal{M} be the set of integers $\{0, \dots, k_n\}$, for any $m \in \mathcal{M}$, we define the model S_m as the space of piecewise polynomials of degree 2 based on the 2^{k_n-m} regular blocks

$$I_i^{(m)} = \bigcup_{j=(i-1)2^m}^{i2^m} I_j, \quad i = 1 \dots, 2^{k_n-m}.$$

We denote by \mathcal{F}^{DP} the collection of models S_m constructed in this way.

PROPOSITION 3.6. *Assume that $(\mathbf{H}_{\text{GauB}})$ or $(\mathbf{H}_{\text{Mom}})$ holds with $p > 6$ in the second case. Let $\eta > 0$ and \tilde{s} be the estimator selected by the procedure (3.1.12) applied to the collection of models \mathcal{F}^{DP} with the penalty*

$$\text{pen}(m) = (1 + \eta) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2. \quad (3.4.1)$$

Suppose also that (\mathbf{A}_3) is fulfilled with the dimension D_n , we define

$$\zeta_n = \frac{\log n - \log D_n}{2 \log D_n} \wedge 1 > 0.$$

For any $\alpha \in (\zeta_n, 1)$ and $L > 0$, the penalized least-squares estimator \tilde{s} satisfies

$$\sup_{(s,t) \in \mathcal{H}_\alpha(L) \times \mathcal{H}_\alpha(L)} \mathbb{E} [\|s - \tilde{s}\|_n^2] \leq C_\alpha n^{-2\alpha/(2\alpha+1)} \quad (3.4.2)$$

where $C_\alpha > 1$ only depends on α , ρ , σ^2 , L , θ and p (under $(\mathbf{H}_{\text{Mom}})$ only).

Note that the supremum is taken over Hölderian balls for the two components of the regression function, *i.e.* the regression function is itself supposed to belong to an Hölderian space. As we mention in the introduction, Stone [Sto85] has proved that the rate of convergence given by (3.4.2) is optimal in the minimax sense. The parameter α belongs to $(\zeta_n, 1)$ that depends on the dimension D_n . For the empirical bound $\delta\sqrt{n}$ with $\delta > 1$, discussed in the previous section, $\zeta_n \leq 1/2$.

3.5. Estimation when σ^2 is unknown

In contrast with Section 3.2, in this section, the variance factor σ^2 is assumed to be unknown in (3.1.7). Since the penalties given by Theorems 3.2 and 3.3 depend on σ^2 , the procedure introduced in the previous sections does not remain available to the statisticians. Thus, we need to estimate σ^2 in order to replace it in the penalty functions. The results of this section give upperbounds for the L^2 risk of the estimators \tilde{s} constructed in such a way.

To estimate the variance factor, we use a residual least-squares estimator $\hat{\sigma}^2$ that we define as follows. Let V be some linear subspace of $\text{Im}(P_n)$ such that

$$\text{Tr}(^t P_n \pi P_n) \leq \text{Tr}(^t P_n P_n)/2 \quad (3.5.1)$$

where π is the orthogonal projection onto V . We define

$$\hat{\sigma}^2 = \frac{n \|Y - \pi Y\|_n^2}{\text{Tr}(^t P_n (I_n - \pi) P_n)}. \quad (3.5.2)$$

First, we assume that the errors are Gaussian. The following result holds.

THEOREM 3.7. *Assume that $(\mathbf{H}_{\text{GauB}})$ holds. For any $\theta > 0$, we define the penalty function*

$$\forall m \in \mathcal{M}, \quad \text{pen}(m) = (1 + \theta) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \hat{\sigma}^2. \quad (3.5.3)$$

Then, for some positive constants C , C' and C'' that only depend on θ , the penalized least-squares estimator \tilde{s} satisfies

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] \leq C \left(\inf_{m \in \mathcal{M}} \mathbb{E} [\|s - \hat{s}_m\|_n^2] + \|s - \pi s\|_n^2 \right) \quad (3.5.4)$$

$$+ \frac{\rho^2(P_n) \sigma^2}{n} \bar{R}_n(\theta) \quad (3.5.5)$$

where we have set

$$\begin{aligned} \bar{R}_n(\theta) = & C' \left[\left(2 + \frac{\|s\|_n^2}{\rho^2(P_n)\sigma^2} \right) \exp \left(-\frac{\theta^2 \text{Tr}(^tP_n P_n)}{32\rho^2(P_n)} \right) \right. \\ & \left. + \sum_{m \in \mathcal{M}} \exp \left(-C'' \frac{\text{Tr}(^tP_n \pi_m P_n)}{\rho^2(P_n)} \right) \right]. \end{aligned}$$

If the errors are only assumed to satisfy moment condition, we have the following theorem.

THEOREM 3.8. *Assume that $(\mathbf{H}_{\text{Mom}})$ holds. Let $\theta > 0$, we consider the penalty function defined by*

$$\forall m \in \mathcal{M}, \text{pen}(m) = (1 + \theta) \frac{\text{Tr}(^tP_n \pi_m P_n)}{n} \hat{\sigma}^2. \quad (3.5.6)$$

For any $0 < q \leq 1$ such that $2(q+1) < p$, the penalized least-squares estimator \tilde{s} satisfies

$$\mathbb{E}[\|s - \tilde{s}\|_n^{2q}]^{1/q} \leq C \left(\inf_{m \in \mathcal{M}} \mathbb{E}[\|s - \hat{s}_m\|_n^2] + 2\|s - \pi s\|_n^2 \right) + \rho^2(P_n)\sigma^2 \bar{R}_n(p, q, \theta)$$

where $C = C(q, \theta)$ and $C' = C'(p, q, \theta)$ are positive constants,

$$\bar{R}_n(p, q, \theta) = \frac{R_n(p, q, \theta)^{1/q}}{n} + C' \tau_p^{1/q} \kappa_n \left(\frac{\|s\|_n^2}{\rho^2(P_n)\sigma^2} + \tau_p \right) \left(\frac{\rho^{2\alpha_p}(P_n)}{\text{Tr}(^tP_n P_n)^{\beta_p}} \right)^{1/q-2/p}$$

with $R_n(p, q, \theta)$ defined as in Theorem 3.3, $(\kappa_n)_{n \in \mathbb{N}} = (\kappa_n(p, q, \theta))_{n \in \mathbb{N}}$ is a sequence of positive numbers that tends to $\kappa = \kappa(p, q, \theta) > 0$ as $\text{Tr}(^tP_n P_n)/\rho^2(P_n)$ tends to infinity and

$$\alpha_p = (p/2 - 1) \vee 1 \text{ and } \beta_p = (p/2 - 1) \wedge 1.$$

Penalties given by (3.5.3) and (3.5.6) are random and allow to construct estimators \tilde{s} when σ^2 is unknown. This approach leads to theoretical upperbounds for the risk of \tilde{s} . Note that we use some generic model V to construct $\hat{\sigma}^2$. This space is quite arbitrary and is pretty much limited to be an half-space of $\text{Im}(P_n)$. The idea is that taking V as some “large” space can lead to a well approximation of the true s and, thus, $Y - \pi Y$ is not far from being centered and its normalized norm is of order σ^2 . However, in practice, it is known that the estimator $\hat{\sigma}^2$ inclined to overestimate the true value of σ^2 as illustrated by Lemmas 3.12 and 3.13. Consequently, the penalty function tends to be larger and the procedure overpenalizes models with high dimension. To offset this phenomenon, a practical solution could be to choose some smaller θ when σ^2 is unknown than when it is known as we discuss at the end of Section 3.6.

3.6. Simulation study

In this section, we study simulations based on the framework given by (3.1.4) with Gaussian errors. First, we introduce the collections of models that we handle and, next, we illustrate the performances of the estimators in practice by many examples.

3.6.1. Collections of models. To perform the simulation study, we consider four collections of models with various complexities. Let us begin with the two collections based on the trigonometric space generated by the functions (\mathbf{T}) . The dimensions D_n and D'_n of E and F , respectively, are positive even integers and such that, for some $\delta > 0$, $D_n = D'_n \leq \delta \sqrt{n/\log n}$. The space E is generated by the vectors

$$\phi_{2k} = (\sin(k\pi x_1), \dots, \sin(k\pi x_n))' \text{ and } \phi_{2k-1} = (\cos(k\pi x_1), \dots, \cos(k\pi x_n))', \quad k = 1, \dots, D_n/2,$$

and F by

$$\psi_{2k} = (\sin(k\pi y_1), \dots, \sin(k\pi y_n))' \text{ and } \psi_{2k-1} = (\cos(k\pi y_1), \dots, \cos(k\pi y_n))', \quad k = 1, \dots, D'_n/2.$$

As we discuss at the end of Section 3.3, the projection P_n onto E along $F + (E + F)^\perp$ admits a spectral norm $\rho(P_n)$ that is supposed to be close to 1 (we give explicit values below). Let \mathcal{M}^T be the

set of integers $\{1, \dots, D_n\}$, for any subset $m \subset \mathcal{M}^T$, the model $S_m \subset E = \text{Im}(P_n)$ is the linear span of $\{\phi_i, i \in m\}$. For any $k \in \{1, \dots, D_n/2\}$, we set

$$m_k^{NT,1} = \{1, \dots, 2k\} \quad , \quad m_k^{NT,2} = \{2i : 1 \leq i \leq k\} \quad , \quad m_k^{NT,3} = \{2i-1 : 1 \leq i \leq k\} \quad ,$$

and we define the collection of models \mathcal{F}^{NT} by

$$\mathcal{F}^{NT} = \left\{ S_{m_k^{NT,i}} : k = 1 \dots, D_n/2, i = 1, 2, 3 \right\} .$$

Note that, for any $i \in \{1, 2, 3\}$ and $k \in \{1, \dots, D_n/2 - 1\}$, $S_{m_k^{NT,i}} \subsetneq S_{m_{k+1}^{NT,i}}$, *i.e.* the models are nested. In particular, it implies that this collection has a small complexity since $N_d \leq 3$, for any $d \in \mathbb{N}$. Concerning our estimation procedure, Corollary 3.4 allows to take some penalty function of the form

$$\text{pen}_{NT}(m) = (1 + C) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \quad (3.6.1)$$

where C is some positive constant. In contrast, we also consider the larger collection

$$\mathcal{F}^{CT} = \{S_m, m \subset \mathcal{M}^T\}$$

that has an higher complexity since $N_d = D_n!/(d!(D_n - d)!)$ is roughly of order $e^{d \log D_n}$. Thus, we need to take a larger penalty as

$$\text{pen}_{CT}(m) = (1 + C + \log D_n) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 . \quad (3.6.2)$$

Similarly, we construct a small and a large collections of models based on the functions (\mathbf{P}_2) as follows. As suggested by Figure 3.2, to get a small spectral norm $\rho(P_n)$, we take $D_n = 3D'_n/2 \leq \delta\sqrt{n}$, for some $\delta > 0$, such that $D_n = 3K$ where $K \in \mathbb{N}$ is the number of blocks of the regular partition of $[-1, 1]$. To construct E and F , we use the three first Legendre polynomials defined on $[-1, 1]$, namely

$$L_0(x) = 1 \quad , \quad L_1(x) = x \quad , \quad L_2(x) = \frac{3x^2 - 1}{2} \quad ,$$

and we define their scaled and translated versions

$$L_j^{(i)}(x) = \begin{cases} L_j(K(x+1) - 2i + 1) & , \text{ if } x \in I_i \\ 0 & , \text{ otherwise} \end{cases} \quad , i = 1, \dots, K, j = 0, 1, 2 \quad ,$$

where the regular blocks are given by

$$I_i = \left[-1 + \frac{2(i-1)}{K}, -1 + \frac{2i}{K} \right] \quad , i = 1, \dots, K .$$

So, the polynomials given by (\mathbf{P}_2) can be written

$$\phi_{3i+j}(x) = L_j^{(i)}(x), \quad i = 1, \dots, K, \quad j = 0, 1, 2$$

and

$$\psi_{2i+j}(x) = L_{j+1}^{(i)}(x), \quad i = 1, \dots, K, \quad j = 0, 1 .$$

Thus, the vectors that generate E and F are given by

$$\phi_k = (\phi_k(x_1), \dots, \phi_k(x_n))', \quad k = 1, \dots, D_n \quad \text{and} \quad \psi_k = (\psi_k(x_1), \dots, \psi_k(y_n))', \quad k = 1, \dots, D'_n .$$

The image set $\text{Im}(P_n)$ is the approximation space of piecewise polynomials of degree no larger than 2 on the regular partition $\Pi = \{I_1, \dots, I_K\}$. Let us denote by m a pair (Π_m, d_m) where Π_m is a partition of $[-1, 1]$ finer than Π and $d_m \in \{0, 1, 2\}$, we define the model S_m as the linear subspace of $\text{Im}(P_n)$ of the piecewise polynomials of degree d_m on the partition Π_m . Note that, a priori, we do not suppose that the partitions Π_m are regular. Let \mathcal{M}^{P_2} be the set of such pairs m and assume that Π is the dyadic regular partition of $[-1, 1]$ with $K = 2^{l_K}$ blocks. We introduce the following elements of \mathcal{M}^{P_2} ,

$$m_l^{NP,i} = \left(\Pi_l^{\text{dyad}}, i \right), \quad l = 0, \dots, l_K, \quad i = 0, 1, 2 ,$$

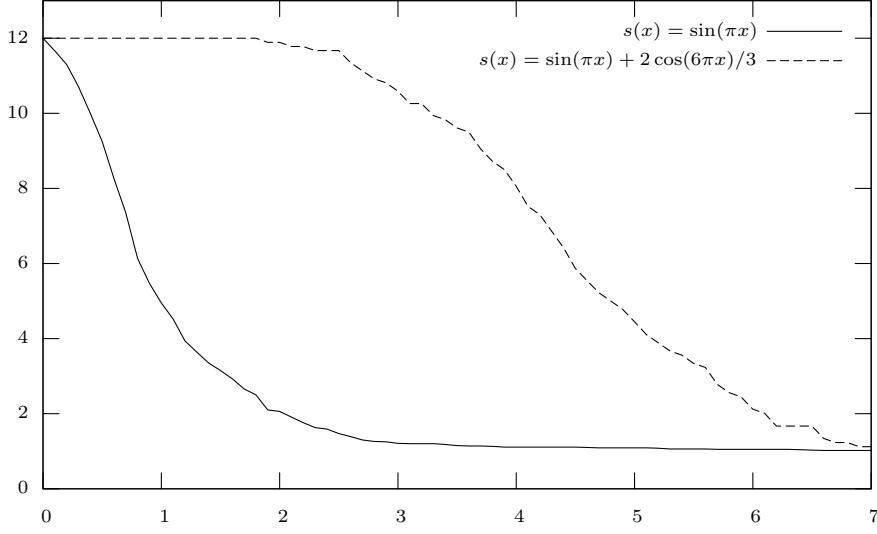


FIGURE 3.3. Estimation of the dimension of $S_{\hat{m}}$ according to the value of C (100 repetitions).

where Π_l^{dyad} is the dyadic regular partition of $[-1, 1]$ with 2^l blocks. Thus, we define the collection of models

$$\mathcal{F}^{NP} = \left\{ S_{m_l^{NP,i}} : l = 0, \dots, l_K, i = 0, 1, 2 \right\}.$$

This collection is nested and have a small complexity since, for any $l \in \{0, \dots, l_K\}$ and $i \in \{0, 1, 2\}$, the dimension of the model $S_{m_l^{NP,i}}$ is $2^l(i+1)$ and, so, $N_d \leq 3$, $d \in \mathbb{N}$. Again, we can consider the penalty given by

$$\text{pen}_{NP}(m) = (1 + C) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2. \quad (3.6.3)$$

If the partition Π is only assumed to be regular with K blocks (and no longer supposed to be a power of 2), we introduce the set $\mathcal{M}_{cons}^{P_2}$ formed by all the elements (Π_m, d_m) of \mathcal{M}^{P_2} such that the blocks of Π_m are consecutive,

$$(\Pi_m, d_m) \in \mathcal{M}_{cons}^{P_2} \iff \text{for all distinct blocks } I, J \in \Pi_m, \sup_{x \in I} x \leq \inf_{y \in J} y \text{ or } \sup_{y \in J} y \leq \inf_{x \in I} x.$$

Finally, we define the collection

$$\mathcal{F}^{CP} = \{ S_m, m \in \mathcal{M}_{cons}^{P_2} \}$$

that has a large complexity. Indeed, for any $d \in \mathbb{N}$, N_d is around $(K-1)!/((d-1)!(K-d)!)$ that is of order $e^{d \log(D_n/3)}$. To deal with this collection, we will consider a penalty function of the form

$$\text{pen}_{CP}(m) = (1 + C + \log(D_n/3)) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2. \quad (3.6.4)$$

3.6.2. The choice of C . To apply our procedure, we need to consider some explicit penalty function. For each collection introduced above, we give a penalty up to some positive constant C that we have to choose in order to proceed. As we mention in the introduction of this chapter, the aim of a model selection procedure is to find some trade-off between a bias term and a variance term. The last one corresponds to the complexity of the model and measures a quantity similar to the dimension. Taking a small value for C will favour the “large” models because the bias term would be more significant than the variance term. In opposite, a large value of C would put the large models at a disadvantage.

Let us illustrate that by considering the estimation of the component s in the framework (3.1.4) with the collection \mathcal{F}^{NT} , $n = 200$, $\delta = 2$, $t(y) = 2y^2 - 2/3$ and a known variance factor $\sigma^2 = 1$. The

figure 3.3 depicts the value of the selected dimension $\dim(S_{\hat{m}})$ when C goes from 0 to 7. When the component s belongs to some small model (namely $S_{m_1^{NT,3}}$ for $s(x) = \sin(\pi x)$), the selected model has a dimension as small as C is large. Conversely, if s is in some model with a large dimension ($s(x) = \sin(\pi x) + 2 \cos(6\pi x)/3$ belongs to $S_{m_6^{NT,1}}$), the procedure will ignore the large models for the benefit of the small ones as C grows.

This phenomenon is well known in model selection theory and, from a theoretical point of view, finding an optimal C is a very hard problem. In order to perform in practice, one usually consider a value slightly larger than 1 for overpenalizing a bit. This choice is empirical and does not claim for optimality. We discuss further about several choices of C .

3.6.3. Numerical simulations. We now illustrate our results and the performances of our estimation procedure by applying it to simulated data

$$Z_i = s(x_i) + t(y_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

where $(x_1, y_1), \dots, (x_n, y_n)$ are known realizations of an uniform random variable on the square $[-1, 1]^2$ and the errors ε_i are i.i.d. standard Gaussian random variables. We handle this framework with known or unknown variance factor σ^2 according to the cases. The unknown functions s and t are chosen among the following ones.

$$\begin{aligned} f_1(t) &= \sin(\pi t) & f_2(t) &= \cos(\pi t) + \cos(3\pi t) & f_3(t) &= \sin(\pi t) + \cos(3\pi t) \\ f_4(t) &= 2t^2 - 2/3 & f_5(t) &= 3 \sin(2\pi(t \vee 0))/2 & f_6(t) &= 2\mathbb{1}_{|t| \leq 0.2} - 0.4 \\ f_7(t) &= 2t & f_8(t) &= 2(\mathbb{1}_{t \leq -1/2} - \mathbb{1}_{-1/2 < t \leq 0}) - \mathbb{1}_{0 < t \leq 1/2} + \mathbb{1}_{1/2 < t} \\ f_9(t) &= e^{2t} - \sinh(2)/2 & f_{10}(t) &= 24(|t| - 3/4)|t| + 1 & f_{11}(t) &= 4(1 + 2t)\mathbb{1}_{0 < t < 1/3} - 8/9 \\ f_{12}(t) &= \begin{cases} 27(t + 1) - 67/9 & , \text{ if } t \leq -1/3 \\ 95/9 + (9t + 3)(2t - 3) & , \text{ if } t > -1/3 \end{cases} \end{aligned}$$

When the variance factor is assumed to be unknown, we estimate it by the procedure described in Section 3.5 and we substitute the quantity σ^2 in the penalties (3.6.1), (3.6.2), (3.6.3) and (3.6.4) by its estimator $\hat{\sigma}^2$. Moreover, in the sequel of this subsection, we take samples of size $n = 200$, a variance $\sigma^2 = 1$ and a factor $\delta = 2$. On all the following figures, the true function s is plotted in dotted line and the estimator \tilde{s} in plain line.

Let us begin with the trigonometric collections of models \mathcal{F}^{NT} and \mathcal{F}^{CT} . In both cases, the second component t is taken equal to the function f_4 . To proceed with \mathcal{F}^{NT} , we use the penalty (3.6.1). A Mallows' heuristic suggests to take a factor equal to two. As we discussed previously, it is usual to slightly overpenalize and we take $C = 1.5$. The figure 3.4 shows the estimator \tilde{s} of the first component s for the choices f_1, f_2, f_3 and f_4 with known and unknown variance factor σ^2 . The estimator performs visually well in the known variance case for functions that belongs to one of the models (f_1, f_2, f_3) or not (f_4). When the variance is unknown, the third case gives a poor estimation of the component s . This particular case illustrates the drawback of estimating σ^2 in some general half-space V . Indeed, it is known that least-squares estimators like $\hat{\sigma}^2$ overestimate the true value of σ^2 . Consequently, the penalty is large and the procedure does not select model of high dimension as $S_{m_3^{NT,1}}$. Only the low-dimensional part of the signal (the sinus term) is detected, the remainder part is considered as noise by the procedure. In the other case, we see that the method does not suffer from the fact that the variance is unknown.

We next handle the collection \mathcal{F}^{CT} with the penalty (3.6.2). Since $\log(D_n) < 2.51$, we take some $C > 0$ such that $1 + C + \log(D_n) = 3.5$. Figure 3.5 illustrates estimations of f_3, f_4, f_5 and f_6 for known and unknown σ^2 . This time, the collection contains a model of adapted dimension to estimate

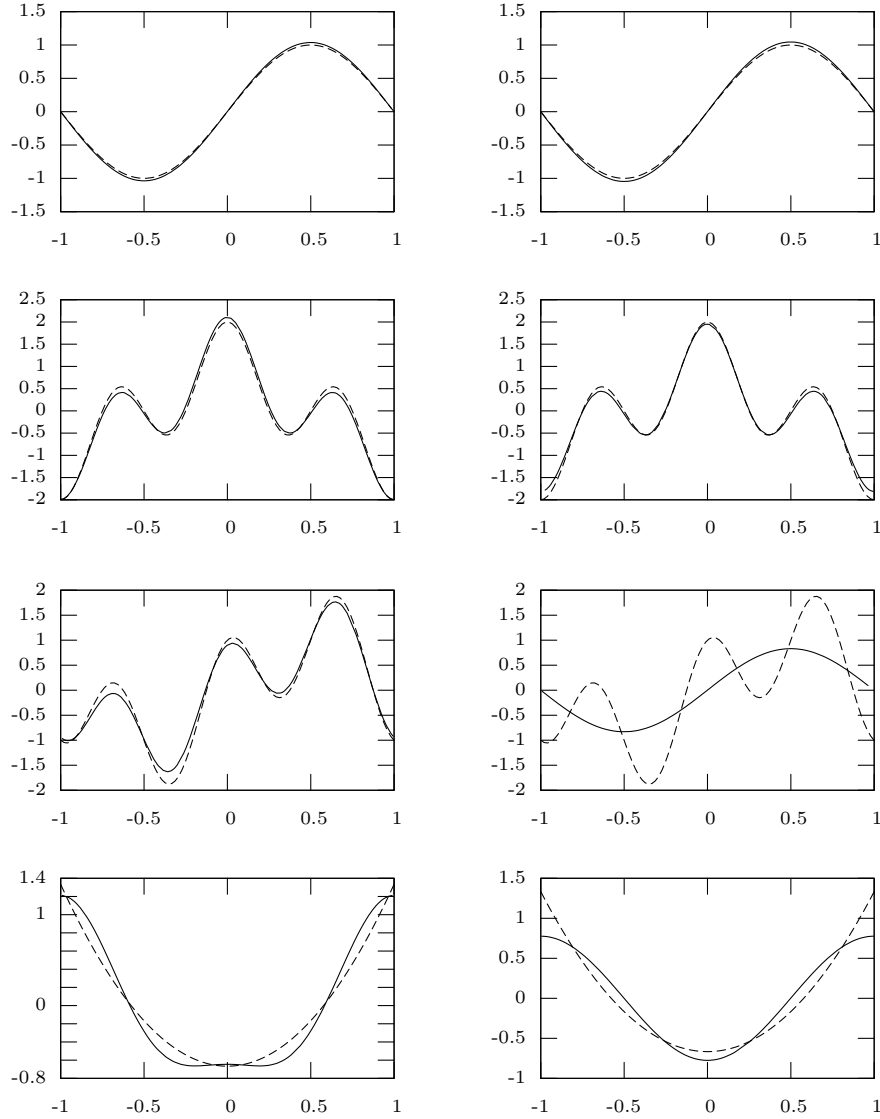


FIGURE 3.4. Estimation of the component $s = f_1, f_2, f_3, f_4$ (from top to bottom) in \mathcal{F}^{NT} with $t = f_4$ for known (left) and unknown (right) variance factor σ^2 (\tilde{s} in plain line, s in dotted line).

f_3 and the procedure select it even if the variance is unknown (*i.e.* the capacity of approximation compensates the overpenalization). As for the nested collection, the quadratic function f_4 is visually well approximated. Moreover, the estimation procedure is able to take into account the changes in the behavior of the component s like in the cases f_5 and f_6 . Another time, we note that the method does not suffer from an unknown variance factor.

We now turn to the piecewise polynomials collections \mathcal{F}^{NP} and \mathcal{F}^{CP} with component $t = f_2$. For the nested one, we consider the penalty (3.6.3) with $C = 1.5$ again and we estimate the first component when it is equal to f_4, f_7, f_8 and f_9 (see Figure 3.6). In the three first situations, we see that we choose the good model for polynomials of degree 0, 1 or 2 for known or unknown σ^2 . For

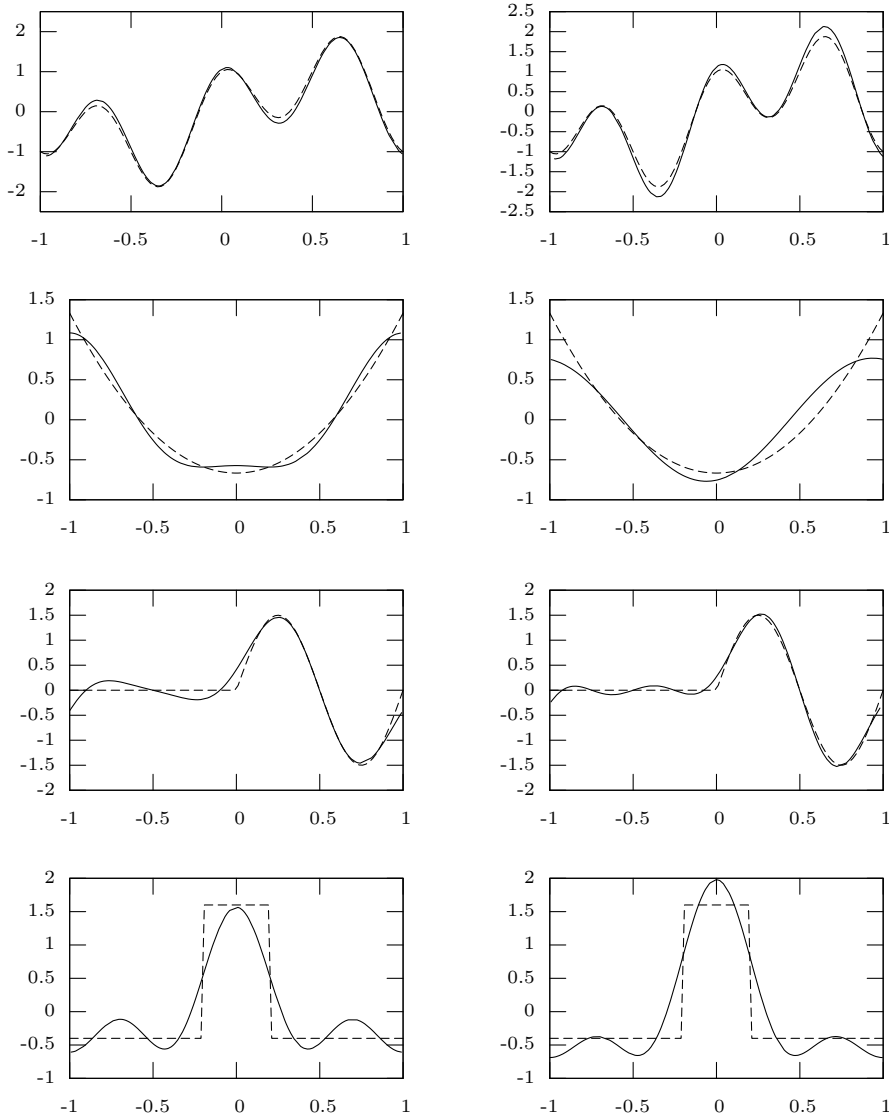


FIGURE 3.5. Estimation of the component $s = f_3, f_4, f_5, f_6$ (from top to bottom) in \mathcal{F}^{CT} with $t = f_4$ for known (left) and unknown (right) variance factor σ^2 (\tilde{s} in plain line, s in dotted line).

some more general function like f_9 , the true function is correctly approximated for known or unknown σ^2 .

Finally, we estimate f_9, f_{10}, f_{11} and f_{12} with the collection \mathcal{F}^{CP} and the penalty (3.6.4). Since the quantity $\log(D_n/3)$ is not larger than 2.24, we take the constant $C > 0$ such that $1 + C + \log(D_n/3) = 3.5$. As with the nested collection, the procedure performs well for f_9 . The discontinuities of the components f_{10}, f_{11} and f_{12} are detected even in the unknown σ^2 case. Again, we can note that the procedure selects models of lower dimension when the variance is unknown. This is due to our previous remark about the tendency of $\hat{\sigma}^2$ to overestimate σ^2 .

3.6.4. Estimation of L^2 ratio. In Section 3.3, we discussed about assumptions that ensure a small remainder term in Inequality (3.3.3). This result corresponds to some oracle type inequality

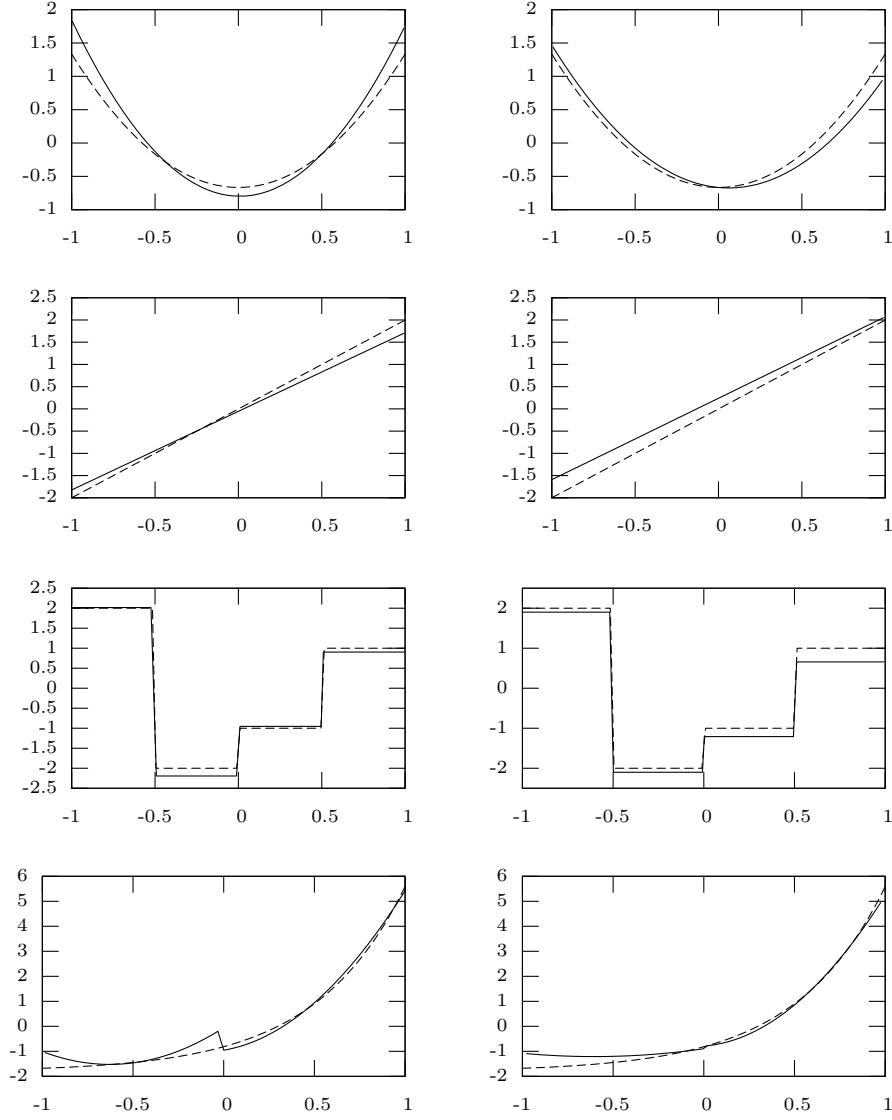


FIGURE 3.6. Estimation of the component $s = f_4, f_7, f_8, f_9$ (from top to bottom) in \mathcal{F}^{NP} with $t = f_2$ for known (left) and unknown (right) variance factor σ^2 (\tilde{s} in plain line, s in dotted line).

for our estimation procedure of a component in an additive framework. Thus, to quantify the performances of our estimator, we are interested in the value of the factor $C(1 + \rho(P_n))^2$. To illustrate that, we estimate the ratio

$$r(\tilde{s}) = \frac{\mathbb{E} [\|s - \tilde{s}\|_n^2]}{\inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \right\}}.$$

We proceed by repeating each of the above simulations 100 times with a penalty function of the form

$$\text{pen}(m) = C \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2$$

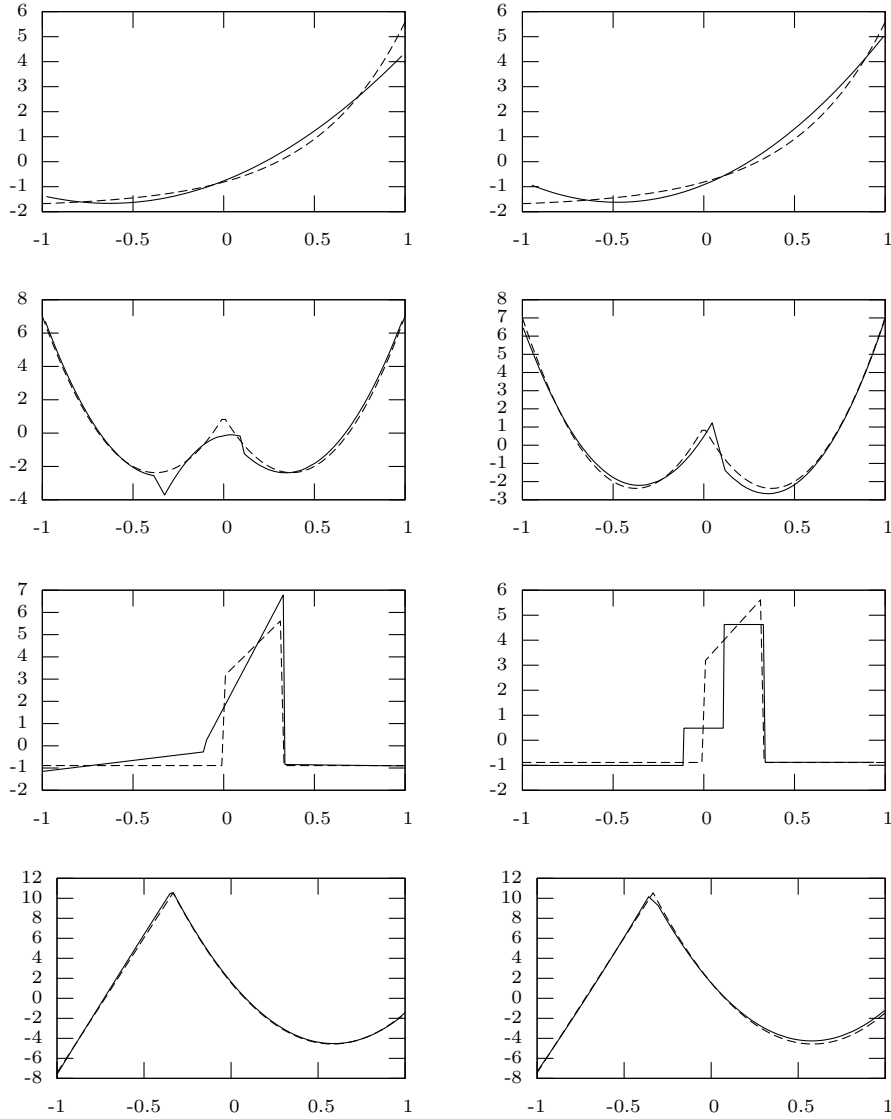


FIGURE 3.7. Estimation of the component $s = f_9, f_{10}, f_{11}, f_{12}$ (from top to bottom) in \mathcal{F}^{CP} with $t = f_2$ for known (left) and unknown (right) variance factor σ^2 (\tilde{s} in plain line, s in dotted line).

with various values of $C > 0$. We estimate $r(\tilde{s})$ by the mean value of the 100 experiments. The obtained results are given for samples of size $n = 200$ and $n = 500$ with known and unknown variance in each situation in Tables 3.1, 3.2, 3.3 and 3.4. When $n = 500$, we restrict our procedure to $\delta = 1.25$ in order to keep the computation time reasonable.

These results show that taking some penalty factor close to 1 is not a good thing, at least in known variance case. When σ^2 is unknown, we recover the phenomenon introduced previously. The values of C that give the small ratios inclined to be smaller than when σ^2 is known. Indeed, to compensate the overpenalization due to large estimation of σ^2 by $\hat{\sigma}^2$, we need to consider some smaller factor C . Moreover, the critical estimation of $s = f_3$ with \mathcal{F}^{NT} and unknown σ^2 is now clear. The half-space V is not large enough to correctly approximated the mean of Y and leads to a (very) large value of $\hat{\sigma}^2$. More generally, we see that the ratios admit order reasonably small for $C \simeq 2.5$ in the nested cases

Parameters	$s(x)$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Known σ^2 $\rho(P_n) = 1.104$ $n = 200$	$f_1(x)$	9.1	5.9	4.0	3.1	2.8	2.0	1.5	1.3	1.2
	$f_2(x)$	2.7	2.1	1.6	1.3	1.2	1.1	1.0	1.0	1.0
	$f_3(x)$	1.5	1.2	1.2	1.1	1.0	1.0	0.9	0.9	0.9
	$f_4(x)$	2.8	2.0	1.8	1.7	1.6	1.7	1.7	1.8	1.7
Unknown σ^2 $\rho(P_n) = 1.104$ $n = 200$	$f_1(x)$	7.4	4.3	3.3	2.8	2.4	2.4	2.3	2.2	2.0
	$f_2(x)$	0.9	0.8	0.8	0.8	1.7	27	30	30	30
	$f_3(x)$	1.0	14	15	15	15	15	15	15	15
	$f_4(x)$	2.1	1.6	1.4	1.4	1.5	1.5	1.6	1.6	1.6
Known σ^2 $\rho(P_n) = 1.069$ $n = 500$	$f_1(x)$	11	7.8	4.7	3.5	2.4	2.3	2.0	2.0	1.8
	$f_2(x)$	3.9	2.8	2.0	1.4	1.2	1.2	1.1	1.0	1.0
	$f_3(x)$	2.3	1.7	1.4	1.2	1.1	1.1	1.1	1.1	1.1
	$f_4(x)$	2.9	2.1	1.9	1.8	1.7	1.5	1.5	1.5	1.6
Unknown σ^2 $\rho(P_n) = 1.069$ $n = 500$	$f_1(x)$	8.1	5.1	3.8	3.3	2.8	2.4	2.2	2.2	2.1
	$f_2(x)$	3.8	2.4	2.0	1.7	1.6	1.4	1.3	1.3	1.2
	$f_3(x)$	2.0	1.4	1.2	1.2	1.1	1.1	1.1	1.1	1.1
	$f_4(x)$	2.4	1.8	1.6	1.5	1.5	1.6	1.6	1.6	1.7

TABLE 3.1. Ratio $r(\tilde{s})$ for the estimation of s with \mathcal{F}^{NT} and $\delta = 2.00$.

Parameters	$s(x)$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Known σ^2 $\rho(P_n) = 1.056$ $n = 200$	$f_3(x)$	2.7	2.4	2.2	2.1	1.8	1.7	1.6	1.4	1.3
	$f_4(x)$	2.0	2.0	1.9	1.9	1.9	2.0	2.0	2.0	2.0
	$f_5(x)$	1.7	1.6	1.5	1.4	1.4	1.3	1.3	1.3	1.3
	$f_6(x)$	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
Unknown σ^2 $\rho(P_n) = 1.056$ $n = 200$	$f_3(x)$	0.9	0.9	1.8	4.8	10	19	38	51	51
	$f_4(x)$	1.7	1.6	1.7	1.7	1.8	1.8	1.8	1.8	1.8
	$f_5(x)$	3.6	4.8	6.4	10	13	15	16	16	16
	$f_6(x)$	1.2	1.8	2.3	2.9	3.2	3.3	3.3	3.3	3.3
Known σ^2 $\rho(P_n) = 1.027$ $n = 500$	$f_3(x)$	3.8	3.3	2.9	2.5	2.2	1.9	1.8	1.7	1.5
	$f_4(x)$	2.5	2.4	2.4	2.3	2.2	2.2	2.1	2.1	2.0
	$f_5(x)$	2.1	2.1	2.0	1.8	1.7	1.6	1.6	1.6	1.5
	$f_6(x)$	1.1	1.1	1.1	1.1	1.1	1.0	1.0	1.1	1.1
Unknown σ^2 $\rho(P_n) = 1.027$ $n = 500$	$f_3(x)$	0.8	0.8	0.8	0.8	0.8	0.8	0.8	3.2	9.3
	$f_4(x)$	2.2	2.0	2.0	1.9	1.8	2.0	1.9	2.0	2.2
	$f_5(x)$	1.3	1.2	1.2	1.1	1.1	1.4	1.7	3.0	4.7
	$f_6(x)$	1.1	1.1	1.2	1.2	1.2	1.3	1.4	1.6	1.9

TABLE 3.2. Ratio $r(\tilde{s})$ for the estimation of s with \mathcal{F}^{CT} and $\delta = 1.25$.

\mathcal{F}^{NT} and \mathcal{F}^{NP} and $C \simeq 3.0$ for \mathcal{F}^{CT} and \mathcal{F}^{CP} when σ^2 is known. In the case of unknown variance factor, taking smaller values for C could be a way for bypassing the overpenalization in practice.

3.7. Proofs

In the proofs, we repeatedly use the following elementary inequality that holds for any $\alpha > 0$ and $x, y \in \mathbb{R}$,

$$2|xy| \leq \alpha x^2 + \alpha^{-1} y^2. \quad (3.7.1)$$

3.7.1. Proofs of Theorems 3.2 and 3.3.

Parameters	$s(x)$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Known σ^2 $\rho(P_n) = 1.223$ $n = 200$	$f_4(x)$	12	9.5	6.0	4.2	3.3	2.5	2.3	2.2	2.1
	$f_7(x)$	16	11	6.8	4.8	4.0	3.1	2.7	2.4	2.4
	$f_8(x)$	8.7	6.7	4.5	3.7	3.2	2.7	2.4	2.0	1.7
	$f_9(x)$	6.0	4.7	3.1	2.4	2.2	2.1	2.2	2.1	2.2
Unknown σ^2 $\rho(P_n) = 1.223$ $n = 200$	$f_4(x)$	6.6	4.4	3.6	3.1	2.6	2.4	2.3	2.2	2.0
	$f_7(x)$	10	5.7	4.3	3.3	2.6	2.3	2.2	2.2	2.1
	$f_8(x)$	6.1	3.8	3.3	2.8	2.4	2.2	1.9	1.7	1.6
	$f_9(x)$	3.7	2.5	2.3	2.2	2.1	2.2	2.2	2.3	2.3
Known σ^2 $\rho(P_n) = 1.252$ $n = 500$	$f_4(x)$	11	8.0	5.4	3.4	2.6	2.4	2.0	1.8	1.8
	$f_7(x)$	16	11	6.1	3.7	3.2	2.0	1.6	1.6	1.4
	$f_8(x)$	8.8	6.5	4.2	2.8	2.4	2.2	1.9	1.7	1.6
	$f_9(x)$	5.2	4.1	3.0	2.3	1.9	1.7	1.7	1.7	1.7
Unknown σ^2 $\rho(P_n) = 1.252$ $n = 500$	$f_4(x)$	6.0	4.3	3.5	3.0	2.5	1.9	1.8	1.7	1.6
	$f_7(x)$	10	5.4	4.1	3.0	2.0	1.8	1.7	1.6	1.5
	$f_8(x)$	4.9	3.3	2.8	2.5	2.1	2.0	1.9	1.8	1.7
	$f_9(x)$	3.2	2.4	2.1	2.1	2.1	2.1	2.2	2.2	2.3

TABLE 3.3. Ratio $r(\tilde{s})$ for the estimation of s with \mathcal{F}^{NP} and $\delta = 2.00$.

Parameters	$s(x)$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Known σ^2 $\rho(P_n) = 1.125$ $n = 200$	$f_9(x)$	4.4	4.1	3.8	3.5	3.3	3.1	2.9	2.9	2.9
	$f_{10}(x)$	2.7	2.6	2.4	2.3	2.1	2.1	2.0	2.0	2.0
	$f_{11}(x)$	1.2	1.2	1.2	1.2	1.1	1.1	1.1	1.2	1.2
	$f_{12}(x)$	2.0	1.9	1.9	1.8	1.7	1.7	1.6	1.5	1.4
Unknown σ^2 $\rho(P_n) = 1.125$ $n = 200$	$f_9(x)$	4.0	3.2	2.9	2.9	2.9	2.9	2.9	3.0	3.0
	$f_{10}(x)$	2.6	2.2	2.1	2.1	2.1	2.1	2.1	2.1	2.3
	$f_{11}(x)$	1.6	2.1	2.1	2.1	2.1	2.7	4.0	5.1	5.4
	$f_{12}(x)$	2.1	2.9	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Known σ^2 $\rho(P_n) = 1.090$ $n = 500$	$f_9(x)$	8.0	7.3	6.5	5.7	5.1	4.9	4.7	4.2	4.1
	$f_{10}(x)$	5.4	5.0	4.7	4.2	3.7	3.4	3.2	3.0	2.9
	$f_{11}(x)$	1.8	1.7	1.7	1.6	1.5	1.4	1.3	1.3	1.3
	$f_{12}(x)$	9.2	8.6	7.6	6.8	6.0	5.1	4.5	4.0	3.6
Unknown σ^2 $\rho(P_n) = 1.090$ $n = 500$	$f_9(x)$	6.0	4.5	3.7	3.4	3.3	3.5	3.6	3.9	4.2
	$f_{10}(x)$	4.4	3.3	2.6	2.3	2.2	2.2	2.2	2.3	2.4
	$f_{11}(x)$	1.2	1.2	1.5	2.0	2.2	2.4	2.4	2.9	4.1
	$f_{12}(x)$	2.9	2.0	1.7	1.6	1.6	1.6	1.6	1.6	1.6

TABLE 3.4. Ratio $r(\tilde{s})$ for the estimation of s with \mathcal{F}^{CP} and $\delta = 1.25$.

Proof of Theorem 3.2. By definition of γ_n , for any $t \in \mathbb{R}^n$, we can write

$$\|s - t\|_n^2 = \gamma_n(t) + 2\sigma \langle t - Y, P_n \varepsilon \rangle_n + \sigma^2 \|P_n \varepsilon\|_n^2.$$

Let $m \in \mathcal{M}$, since $\hat{s}_m = s_m + \sigma \pi_m P_n \varepsilon$, this identity and (3.1.13) lead to

$$\begin{aligned} \|s - \tilde{s}\|_n^2 &= \|s - s_m\|_n^2 + \gamma_n(\tilde{s}) - \gamma_n(s_m) + 2\sigma \langle \tilde{s} - s_m, P_n \varepsilon \rangle_n \\ &= \|s - s_m\|_n^2 + \gamma_n(\tilde{s}) - \gamma_n(\hat{s}_m) - \sigma^2 \|\pi_m P_n \varepsilon\|_n^2 \\ &\quad - 2\sigma \langle s - \tilde{s}, P_n \varepsilon \rangle_n + 2\sigma \langle s - s_m, P_n \varepsilon \rangle_n \\ &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{pen}(\hat{m}) + 2\sigma^2 \|\pi_{\hat{m}} P_n \varepsilon\|_n^2 \\ &\quad - 2\sigma \langle s - s_{\hat{m}}, P_n \varepsilon \rangle_n + 2\sigma \langle s - s_m, P_n \varepsilon \rangle_n - \sigma^2 \|\pi_m P_n \varepsilon\|_n^2. \end{aligned} \quad (3.7.2)$$

Consider an arbitrary $a_m \in S_m^\perp$ such that $\|a_m\|_n = 1$, we define

$$u_m = \begin{cases} (s - s_m)/\|s - s_m\|_n & \text{if } s \neq \pi_m s \\ a_m & \text{otherwise.} \end{cases} \quad (3.7.3)$$

Thus, (3.7.2) gives

$$\begin{aligned} \|s - \tilde{s}\|_n^2 &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{pen}(\hat{m}) + 2\sigma^2 \|\pi_{\hat{m}} P_n \varepsilon\|_n^2 \\ &\quad + 2\sigma \|s - s_{\hat{m}}\|_n |\langle u_{\hat{m}}, P_n \varepsilon \rangle_n| + 2\sigma \langle s - s_m, P_n \varepsilon \rangle_n - \sigma^2 \|\pi_m P_n \varepsilon\|_n^2. \end{aligned} \quad (3.7.4)$$

Take $\alpha \in (0, 1)$ that we precise later and we use the inequality (3.7.1),

$$\begin{aligned} (1 - \alpha) \|s - \tilde{s}\|_n^2 &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{pen}(\hat{m}) + (2 - \alpha) \sigma^2 \|\pi_{\hat{m}} P_n \varepsilon\|_n^2 \\ &\quad + \alpha^{-1} \sigma^2 \langle u_{\hat{m}}, P_n \varepsilon \rangle_n^2 + 2\sigma \langle s - s_m, P_n \varepsilon \rangle_n - \sigma^2 \|\pi_m P_n \varepsilon\|_n^2. \end{aligned} \quad (3.7.5)$$

We choose $\alpha = 1/(1 + \theta) \in (0, 1)$ but for legibility we keep using the notation α . Let us now introduce two functions $p_1, p_2 : \mathcal{M} \rightarrow \mathbb{R}_+$ that will be specified later to satisfy, for all $m \in \mathcal{M}$,

$$\text{pen}(m) \geq (2 - \alpha) p_1(m) + \alpha^{-1} p_2(m). \quad (3.7.6)$$

We use this bound in (3.7.5) to obtain

$$\begin{aligned} (1 - \alpha) \|s - \tilde{s}\|_n^2 &\leq \|s - s_m\|_n^2 + \text{pen}(m) + (2 - \alpha) (\sigma^2 \|\pi_{\hat{m}} P_n \varepsilon\|_n^2 - p_1(\hat{m})) \\ &\quad + \alpha^{-1} (\sigma^2 \langle u_{\hat{m}}, P_n \varepsilon \rangle_n^2 - p_2(\hat{m})) + 2\sigma \langle s - s_m, P_n \varepsilon \rangle_n \\ &\quad - \sigma^2 \|\pi_m P_n \varepsilon\|_n^2 \\ &\leq \|s - s_m\|_n^2 + \text{pen}(m) + 2\sigma \langle s - s_m, P_n \varepsilon \rangle_n - \sigma^2 \|\pi_m P_n \varepsilon\|_n^2 \\ &\quad + (2 - \alpha) \sup_{m' \in \mathcal{M}} (\sigma^2 \|\pi_{m'} P_n \varepsilon\|_n^2 - p_1(m'))_+ \\ &\quad + \alpha^{-1} \sup_{m' \in \mathcal{M}} (\sigma^2 \langle u_{m'}, P_n \varepsilon \rangle_n^2 - p_2(m'))_+. \end{aligned}$$

Taking the expectation on both side, it leads to

$$\begin{aligned} (1 - \alpha) \mathbb{E} [\|s - \tilde{s}\|_n^2] &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{Tr}(^t P_n \pi_m P_n) \sigma^2 / n \\ &\quad + (2 - \alpha) \mathbb{E} \left[\sup_{m' \in \mathcal{M}} (\sigma^2 \|\pi_{m'} P_n \varepsilon\|_n^2 - p_1(m'))_+ \right] \\ &\quad + \alpha^{-1} \mathbb{E} \left[\sup_{m' \in \mathcal{M}} (\sigma^2 \langle u_{m'}, P_n \varepsilon \rangle_n^2 - p_2(m'))_+ \right] \\ &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{Tr}(^t P_n \pi_m P_n) \sigma^2 / n \\ &\quad + (2 - \alpha) \sum_{m' \in \mathcal{M}} \mathbb{E} \left[(\sigma^2 \|\pi_{m'} P_n \varepsilon\|_n^2 - p_1(m'))_+ \right] \\ &\quad + \alpha^{-1} \sum_{m' \in \mathcal{M}} \mathbb{E} \left[(\sigma^2 \langle u_{m'}, P_n \varepsilon \rangle_n^2 - p_2(m'))_+ \right] \\ &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{Tr}(^t P_n \pi_m P_n) \sigma^2 / n \\ &\quad + (2 - \alpha) \sum_{m' \in \mathcal{M}} \mathbb{E}_{1, m'} + \alpha^{-1} \sum_{m' \in \mathcal{M}} \mathbb{E}_{2, m'}. \end{aligned}$$

Because the choice of m is arbitrary among \mathcal{M} , we can infer that

$$(1 - \alpha) \mathbb{E} [\|s - \tilde{s}\|_n^2] \leq \inf_{m \in \mathcal{M}} \{ \|s - s_m\|_n^2 + \text{pen}(m) - \text{Tr}(^t P_n \pi_m P_n) \sigma^2 / n \} \quad (3.7.7)$$

$$+ (2 - \alpha) \sum_{m \in \mathcal{M}} \mathbb{E}_{1,m} + \alpha^{-1} \sum_{m \in \mathcal{M}} \mathbb{E}_{2,m}.$$

We now have to upperbound $\mathbb{E}_{1,m}$ and $\mathbb{E}_{2,m}$ in (3.7.7). Let start by the first one. If $S_m = \{0\}$, then $\pi_m P_n = 0$ and $p_1(m) \geq 0$ suffices to ensure that $\mathbb{E}_{1,m} = 0$. So, we can consider that the dimension of S_m is positive and $\pi_m P_n \neq 0$. The Lemma 3.10 applied with $A = \pi_m P_n$ gives, for any $x > 0$,

$$\mathbb{P} \left(n \|\pi_m P_n \varepsilon\|_n^2 \geq \text{Tr}(^t P_n \pi_m P_n) + 2\sqrt{\rho^2(P_n) \text{Tr}(^t P_n \pi_m P_n) x} + \rho^2(P_n) x \right) \leq e^{-x} \quad (3.7.8)$$

because $\rho(\pi_m P_n) \leq \rho(\pi_m) \rho(P_n) \leq \rho(P_n)$. Let $\beta = \theta^2 / (1 + 2\theta) > 0$, (3.7.1) and (3.7.8) lead to

$$\mathbb{P} \left(n \|\pi_m P_n \varepsilon\|_n^2 \geq (1 + \beta) \text{Tr}(^t P_n \pi_m P_n) + (1 + \beta^{-1}) \rho^2(P_n) x \right) \leq e^{-x}. \quad (3.7.9)$$

Let $\delta = \theta^2 / ((1 + \theta)(1 + 2\theta + 2\theta^2)) > 0$, we set

$$np_1(m) = ((1 + \beta) + (1 + \beta^{-1}) \delta L_m) \text{Tr}(^t P_n \pi_m P_n) \sigma^2$$

and (3.7.9) implies

$$\begin{aligned} \mathbb{E}_{m,1} &= \int_0^\infty \mathbb{P} \left((\sigma^2 \|\pi_m P_n \varepsilon\|_n^2 - p_1(m))_+ \geq \xi \right) d\xi \\ &= \int_0^\infty \mathbb{P} \left(n \|\pi_m P_n \varepsilon\|_n^2 - np_1(m) / \sigma^2 \geq n\xi / \sigma^2 \right) d\xi \\ &\leq \int_0^\infty \exp \left(-\frac{\delta L_m \text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} - \frac{n\xi}{(1 + \beta^{-1}) \rho^2(P_n) \sigma^2} \right) d\xi \\ &\leq \frac{(1 + \beta^{-1}) \rho^2(P_n) \sigma^2}{n} \exp \left(-\frac{\delta L_m \text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} \right). \end{aligned} \quad (3.7.10)$$

We now focus on $\mathbb{E}_{m,2}$. The random variable $\langle u_m, P_n \varepsilon \rangle_n = \langle ^t P_n u_m, \varepsilon \rangle_n$ is a centered Gaussian variable with variance $\| ^t P_n u_m \|_n^2 / n$. For any $x > 0$, the standard Gaussian deviation inequality gives

$$\mathbb{P} (|\langle u_m, P_n \varepsilon \rangle_n| \geq x) \leq \exp \left(-\frac{nx^2}{2 \| ^t P_n u_m \|_n^2} \right) \leq \exp \left(-\frac{nx^2}{2 \rho^2(P_n)} \right)$$

that is equivalent to

$$\mathbb{P} (n \langle u_m, P_n \varepsilon \rangle_n^2 \geq 2 \rho^2(P_n) x) \leq e^{-x}. \quad (3.7.11)$$

We set

$$np_2(m) = 2 \delta L_m \text{Tr}(^t P_n \pi_m P_n) \sigma^2$$

and (3.7.11) leads to

$$\begin{aligned} \mathbb{E}_{m,2} &= \int_0^\infty \mathbb{P} \left((\sigma^2 \langle u_m, P_n \varepsilon \rangle_n^2 - p_2(m))_+ \geq \xi \right) d\xi \\ &= \int_0^\infty \mathbb{P} \left(\langle u_m, P_n \varepsilon \rangle_n^2 - np_2(m) / \sigma^2 \geq n\xi / \sigma^2 \right) d\xi \\ &\leq \int_0^\infty \exp \left(-\frac{\delta L_m \text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} - \frac{n\xi}{2 \rho^2(P_n) \sigma^2} \right) d\xi \\ &\leq \frac{2 \rho^2(P_n) \sigma^2}{n} \exp \left(-\frac{\delta L_m \text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} \right). \end{aligned} \quad (3.7.12)$$

We inject (3.7.10) and (3.7.12) in (3.7.7) and we replace α , β and δ to obtain

$$\begin{aligned} \frac{\theta}{\theta + 1} \mathbb{E} [\|s - \tilde{s}\|_n^2] &\leq \inf_{m \in \mathcal{M}} \{ \|s - s_m\|_n^2 + \text{pen}(m) - \text{Tr}(^t P_n \pi_m P_n) \sigma^2 / n \} \\ &\quad + C_\theta \frac{\rho^2(P_n) \sigma^2}{n} R_\theta \end{aligned}$$

where we have set

$$\begin{aligned} R_\theta &= \sum_{m \in \mathcal{M}} \exp \left(-\frac{\theta^2 L_m \text{Tr}(^t P_n \pi_m P_n)}{(1+\theta)(1+2\theta+2\theta^2)\rho^2(P_n)} \right) \\ &\leq \sum_{m \in \mathcal{M}} \exp \left(-\frac{\theta^2 L_m \text{Tr}(^t P_n \pi_m P_n)}{2(1+\theta)^3 \rho^2(P_n)} \right) \end{aligned}$$

and

$$C_\theta = \frac{(1+\theta)(1+2\theta+2\theta^2)}{\theta^2} \leq \frac{2(1+\theta)^3}{\theta^2}.$$

Finally, (3.7.6) gives a penalty as (3.2.1) and the announced result follows.

Proof of Theorem 3.3. In order to prove Theorem 3.3, we show the following stronger result. Under the assumptions of the theorem, there exists a positive constant C that only depends on p and θ , such that, for any $z > 0$,

$$\mathbb{P} \left(\frac{\theta}{\theta+2} \mathcal{H}_+ \geq \frac{\rho^2(P_n) \sigma^2}{n} z \right) \leq C \tau_p \left[N_0 \left(1 \wedge z^{-p/2} \right) + R_{P_n, p}(\mathcal{F}, z) \right] \quad (3.7.13)$$

where the quantity \mathcal{H} is defined by

$$\mathcal{H} = \|s - \tilde{s}\|_n^2 - \frac{\theta+4}{\theta} \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{2(\theta+2)}{\theta+4} \text{pen}(m) \right\}$$

and we have set

$$R_{P_n, p}(\mathcal{F}, z) = \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \left(1 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho(^t P_n \pi_m P_n)} \right) \left(\frac{L_m \text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} + z \right)^{-p/2}.$$

Thus, for any $q > 0$ such that $2(q+1) < p$, we integrate (3.7.13) via Lemma 3.9 to get

$$\begin{aligned} \mathbb{E} [\mathcal{H}_+^q] &= \int_0^\infty q t^{q-1} \mathbb{P}(\mathcal{H}_+ \geq t) dt \\ &= \left(\frac{(\theta+2)\rho^2(P_n)\sigma^2}{\theta n} \right)^q \int_0^\infty q z^{q-1} \mathbb{P} \left(\frac{\theta}{\theta+2} \mathcal{H}_+ \geq \frac{\rho^2(P_n)\sigma^2}{n} z \right) dz \\ &\leq C'(p, q, \theta) \tau_p \left(\frac{\rho^2(P_n)\sigma^2}{n} \right)^q R_{P_n, \theta}^{p, q}(\mathcal{F}) \end{aligned} \quad (3.7.14)$$

where we have set

$$R_{P_n, \theta}^{p, q}(\mathcal{F}) = N_0 + \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \left(1 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho(^t P_n \pi_m P_n)} \right) \left(\frac{L_m \text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} \right)^{q-p/2}.$$

Since

$$\mathbb{E} [\|s - \tilde{s}\|_n^{2q}]^{1/q} \leq \mathbb{E} \left[\left(\frac{\theta+8}{\theta} \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{2(\theta+4)}{\theta+8} \text{pen}(m) \right\} + \mathcal{H}_+ \right)^q \right]^{1/q},$$

it follows from Minkowski's Inequality when $q \geq 1$ or convexity arguments when $0 < q < 1$ that

$$\begin{aligned} \mathbb{E} [\|s - \tilde{s}\|_n^{2q}]^{1/q} &\leq 2^{(q-1)_+} \left(C''(\theta) \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \text{pen}(m) \right\} + \mathbb{E} [\mathcal{H}_+^q]^{1/q} \right). \end{aligned} \quad (3.7.15)$$

Inequality (3.2.4) directly follows from (3.7.14) and (3.7.15).

We now turn to the proof of (3.7.13). Inequality (3.7.5) does not depend on the distribution of ε and we start from here. Let $\alpha = \alpha(\theta) \in (0, 1)$, for any $m \in \mathcal{M}$ we have

$$\begin{aligned} (1-\alpha)\|s - \tilde{s}\|_n^2 &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{pen}(\hat{m}) + (2-\alpha)\sigma^2 \|\pi_{\hat{m}} P_n \varepsilon\|_n^2 \\ &\quad + \alpha^{-1} \sigma^2 \langle u_{\hat{m}}, P_n \varepsilon \rangle_n^2 + 2\sigma \langle s - s_m, P_n \varepsilon \rangle_n \end{aligned}$$

where u_m is defined by (3.7.3). Use again (3.7.1) with α to obtain

$$\begin{aligned} (1 - \alpha)\|s - \tilde{s}\|_n^2 &\leq \|s - s_m\|_n^2 + \text{pen}(m) - \text{pen}(\hat{m}) + (2 - \alpha)\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2 \\ &\quad + \alpha^{-1}\sigma^2\langle u_{\hat{m}}, P_n\varepsilon \rangle_n^2 + 2\sigma\|s - s_m\|_n|\langle u_m, P_n\varepsilon \rangle_n| \\ &\leq (1 + \alpha)\|s - s_m\|_n^2 + \text{pen}(m) - \text{pen}(\hat{m}) + (2 - \alpha)\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2 \\ &\quad + \alpha^{-1}\sigma^2\langle u_{\hat{m}}, P_n\varepsilon \rangle_n^2 + \alpha^{-1}\sigma^2\langle u_m, P_n\varepsilon \rangle_n^2. \end{aligned} \quad (3.7.16)$$

Let us now introduce two functions $\bar{p}_1, \bar{p}_2 : \mathcal{M} \rightarrow \mathbb{R}_+$ that will be specified later and that satisfy,

$$\forall m \in \mathcal{M}, \text{pen}(m) \geq (2 - \alpha)\bar{p}_1(m) + \alpha^{-1}\bar{p}_2(m). \quad (3.7.17)$$

Thus, Inequality (3.7.16) implies

$$\begin{aligned} (1 - \alpha)\|s - \tilde{s}\|_n^2 &\leq (1 + \alpha)\|s - s_m\|_n^2 + \text{pen}(m) + \alpha^{-1}\bar{p}_2(m) \\ &\quad + (2 - \alpha)(\sigma^2\|\pi_{\hat{m}}P_n\varepsilon\|_n^2 - \bar{p}_1(\hat{m})) \\ &\quad + \alpha^{-1}(\sigma^2\langle u_{\hat{m}}, P_n\varepsilon \rangle_n^2 - \bar{p}_2(\hat{m})) \\ &\quad + \alpha^{-1}(\sigma^2\langle u_m, P_n\varepsilon \rangle_n^2 - \bar{p}_2(m)) \\ &\leq (1 + \alpha)(\|s - s_m\|_n^2 + 2\text{pen}(m)/(1 + \alpha)) \\ &\quad + (2 - \alpha) \sup_{m' \in \mathcal{M}} (\sigma^2\|\pi_{m'}P_n\varepsilon\|_n^2 - \bar{p}_1(m'))_+ \\ &\quad + 2\alpha^{-1} \sup_{m' \in \mathcal{M}} (\sigma^2\langle u_{m'}, P_n\varepsilon \rangle_n^2 - \bar{p}_2(m'))_+. \end{aligned}$$

Because the choice of m is arbitrary among \mathcal{M} , we can infer that, for any $\xi > 0$,

$$\begin{aligned} \mathbb{P}((1 - \alpha)\mathcal{H}_+ \geq \xi) &\leq \mathbb{P}\left((2 - \alpha) \sup_{m \in \mathcal{M}} (\sigma^2\|\pi_m P_n\varepsilon\|_n^2 - \bar{p}_1(m))_+ \geq \frac{\xi}{2}\right) \\ &\quad + \mathbb{P}\left(2\alpha^{-1} \sup_{m \in \mathcal{M}} (\sigma^2\langle u_m, P_n\varepsilon \rangle_n^2 - \bar{p}_2(m))_+ \geq \frac{\xi}{2}\right) \\ &\leq \sum_{m \in \mathcal{M}} \mathbb{P}\left(\sigma^2\|\pi_m P_n\varepsilon\|_n^2 \geq \bar{p}_1(m) + \frac{\xi}{2(2 - \alpha)}\right) \\ &\quad + \sum_{m \in \mathcal{M}} \mathbb{P}\left(\sigma^2\langle u_m, P_n\varepsilon \rangle_n^2 \geq \bar{p}_2(m) + \frac{\alpha\xi}{4}\right) \\ &\leq \sum_{m \in \mathcal{M}} \mathbb{P}_{1,m}(\xi) + \sum_{m \in \mathcal{M}} \mathbb{P}_{2,m}(\xi). \end{aligned} \quad (3.7.18)$$

We first bound $\mathbb{P}_{1,m}(\xi)$. For $m \in \mathcal{M}$ such that $S_m = \{0\}$ (i.e. $\pi_m = 0$), $\bar{p}_1(m) \geq 0$ leads obviously to $\mathbb{P}_{1,m}(\xi) = 0$. Thus, it is sufficient to bound $\mathbb{P}_{1,m}(\xi)$ for m such that π_m is not the null matrix. This ensures that the symmetric nonnegative matrix $\tilde{A} = {}^tP_n\pi_m P_n$ lies in $\mathbb{M}_n \setminus \{0\}$. Thus, under hypothesis (3.1.6), Corollary 5.1 of [Bar00] gives us, for any $x_m > 0$,

$$\mathbb{P}\left(n\|\pi_m P_n\varepsilon\|_n^2 \geq \text{Tr}(\tilde{A}) + 2\sqrt{\rho(\tilde{A})\text{Tr}(\tilde{A})x_m} + \rho(\tilde{A})x_m\right) \leq \frac{C_1(p)\tau_p\text{Tr}(\tilde{A})}{\rho(\tilde{A})x_m^{p/2}}$$

where $C_1(p)$ is a constant that only depends on p . The properties of the norm ρ imply

$$\rho(\tilde{A}) = \rho({}^t(\pi_m P_n)(\pi_m P_n)) = \rho(\pi_m P_n)^2 \leq \rho^2(P_n). \quad (3.7.19)$$

By the inequalities (3.7.19) and (3.7.1) with $\theta/2 > 0$, we obtain

$$\mathbb{P}\left(n\|\pi_m P_n\varepsilon\|_n^2 \geq \left(1 + \frac{\theta}{2}\right)\text{Tr}(\tilde{A}) + \left(1 + \frac{2}{\theta}\right)\rho^2(P_n)x_m\right) \leq \frac{C_1(p)\tau_p\text{Tr}(\tilde{A})}{\rho(\tilde{A})x_m^{p/2}}. \quad (3.7.20)$$

We take $\alpha = 2/(\theta + 2) \in (0, 1)$ but for legibility we keep using the notation α . Moreover, we choose

$$n\bar{p}_1(m) = \left(1 + \frac{\theta}{2} + \frac{L_m}{2(\theta + 1)}\right)\text{Tr}({}^tP_n\pi_m P_n)\sigma^2$$

and

$$x_m = \frac{\theta}{2(\theta+1)(\theta+2)} \times \frac{L_m \text{Tr}(^t P_n \pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)}.$$

Thus, Inequality (3.7.20) leads to

$$\begin{aligned} \mathbb{P}_{1,m}(\xi) &= \mathbb{P}\left(\sigma^2 \|\pi_m P_n \varepsilon\|_n^2 \geq \bar{p}_1(m) + \frac{\xi}{2(2-\alpha)}\right) \\ &= \mathbb{P}\left(\sigma^2 \|\pi_m P_n \varepsilon\|_n^2 \geq \bar{p}_1(m) + \frac{(\theta+2)\xi}{4(\theta+1)}\right) \\ &\leq \mathbb{P}\left(n \|\pi_m P_n \varepsilon\|_n^2 \geq \left(1 + \frac{\theta}{2}\right) \text{Tr}(^t P_n \pi_m P_n) + \left(1 + \frac{2}{\theta}\right) \rho^2(P_n) x_m\right) \\ &\leq C_2(p, \theta) \frac{\text{Tr}(^t P_n \pi_m P_n) \tau_p}{\rho(^t P_n \pi_m P_n)} \left(\frac{L_m \text{Tr}(^t P_n \pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)}\right)^{-p/2}. \end{aligned} \quad (3.7.21)$$

We now focus on $\mathbb{P}_{2,m}(\xi)$. Let y_m be some positive real number, the Markov Inequality leads to

$$\mathbb{P}(|\langle u_m, P_n \varepsilon \rangle_n| \geq y_m) \leq y_m^{-p} \mathbb{E}[|\langle u_m, P_n \varepsilon \rangle_n|^p] = y_m^{-p} \mathbb{E}[|\langle ^t P_n u_m, \varepsilon \rangle_n|^p]. \quad (3.7.22)$$

Since $p > 2$, the quantity τ_p is lower bounded by 1,

$$\tau_p = \mathbb{E}[|\varepsilon_1|^p] \geq \mathbb{E}[\varepsilon_1^2]^{p/2} = 1. \quad (3.7.23)$$

Moreover, we can apply the Rosenthal inequality (see Chapter 2 of [Pet95]) to obtain

$$\mathbb{E}[|\langle ^t P_n u_m, \varepsilon \rangle_n|^p] \leq C_3(p) n^{-p} \left(\tau_p \sum_{i=1}^n |(^t P_n u_m)_i|^p + n^{p/2} \|^t P_n u_m\|_n^p \right) \quad (3.7.24)$$

where $C_3(p)$ is a constant that only depends on p . Since $p > 2$, we have

$$\sum_{i=1}^n |(^t P_n u_m)_i|^p \leq \left(\sum_{i=1}^n (^t P_n u_m)_i^2 \right)^{p/2} = n^{p/2} \|^t P_n u_m\|_n^p \leq n^{p/2} \rho^p(P_n).$$

Thus, the Inequality (3.7.24) becomes

$$\mathbb{E}[|\langle ^t P_n u_m, \varepsilon \rangle_n|^p] \leq 2C_3(p) \rho^p(P_n) \tau_p n^{-p/2}$$

and, putting this inequality in (3.7.22), we obtain

$$\mathbb{P}(|\langle u_m, P_n \varepsilon \rangle_n| \geq y_m) \leq 2C_3(p) \rho^p(P_n) \tau_p n^{-p/2} y_m^{-p}. \quad (3.7.25)$$

We take

$$n\bar{p}_2(m) = \frac{1}{2(\theta+1)} \sigma^2 L_m \text{Tr}(^t P_n \pi_m P_n)$$

and

$$y_m^2 = \frac{1}{2(\theta+2)n} \left(L_m \text{Tr}(^t P_n \pi_m P_n) + \frac{n\xi}{\sigma^2} \right).$$

Finally, (3.7.25) gives

$$\begin{aligned} \mathbb{P}_{2,m}(\xi) &= \mathbb{P}\left(\sigma^2 \langle u_m, P_n \varepsilon \rangle_n^2 \geq \bar{p}_2(m) + \frac{\alpha\xi}{4}\right) \\ &= \mathbb{P}\left(\sigma^2 \langle u_m, P_n \varepsilon \rangle_n^2 \geq \bar{p}_2(m) + \frac{\xi}{2(\theta+2)}\right) \\ &\leq \mathbb{P}\left(\langle u_m, P_n \varepsilon \rangle_n^2 \geq y_m^2\right) \\ &\leq C_4(p, \theta) \tau_p \left(\frac{L_m \text{Tr}(^t P_n \pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)}\right)^{-p/2}. \end{aligned} \quad (3.7.26)$$

Putting together Inequalities (3.7.18), (3.7.21) and (3.7.26) leads us to

$$\begin{aligned}
& \mathbb{P}((1-\alpha)\mathcal{H}_+ \geq \xi) \\
& \leq \sum_{m \in \mathcal{M}} \mathbb{P}_{1,m}(\xi) + \sum_{m \in \mathcal{M}} \mathbb{P}_{2,m}(\xi) \\
& \leq \sum_{m \in \mathcal{M}: S_m = \{0\}} \mathbb{P}_{2,m}(\xi) + \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \mathbb{P}_{1,m}(\xi) + \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \mathbb{P}_{2,m}(\xi) \\
& \leq \sum_{m \in \mathcal{M}: S_m = \{0\}} 1 \wedge \left\{ C_4(p, \theta) \tau_p \left(\frac{n\xi}{\sigma^2 \rho^2(P_n)} \right)^{-p/2} \right\} \\
& \quad + C_5(p, \theta) \tau_p \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \left(1 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho(^t P_n \pi_m P_n)} \right) \left(\frac{L_m \text{Tr}(^t P_n \pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)} \right)^{-p/2} \\
& \leq N_0(1 \vee C_4(p\theta)) \tau_p \left(1 \wedge \left(\frac{n\xi}{\rho^2(P_n) \sigma^2} \right)^{-p/2} \right) \\
& \quad + C_5(p, \theta) \tau_p \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \left(1 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho(^t P_n \pi_m P_n)} \right) \left(\frac{L_m \text{Tr}(^t P_n \pi_m P_n) + n\xi/\sigma^2}{\rho^2(P_n)} \right)^{-p/2}.
\end{aligned}$$

For $z > 0$, take $\xi = \rho^2(P_n) \sigma^2 z / n$ to obtain (3.7.13). We conclude the proof by computing the lowerbound (3.7.17) on the penalty function,

$$\begin{aligned}
(2-\alpha)\bar{p}_1(m) + \alpha^{-1}\bar{p}_2(m) &= \frac{2(\theta+1)}{\theta+2} \bar{p}_1(m) + \frac{\theta+2}{2} \bar{p}_2(m) \\
&= \left(1 + \theta + \frac{\theta^2 + 8\theta + 8}{4(\theta+1)(\theta+2)} L_m \right) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2.
\end{aligned}$$

Since $(\theta^2 + 8\theta + 8)/(4(\theta+1)(\theta+2)) \leq 1$, the penalty given by (3.2.3) satisfies the condition (3.7.17).

3.7.2. Proofs of Theorems 3.7 and 3.8.

Proof of Theorem 3.7. Given $\theta > 0$, we can find two positive numbers $\delta = \delta(\theta) < 1/2$ and $\eta = \eta(\theta)$ such that $(1+\theta)(1-2\delta) \geq (1+2\eta)$. Thus we define

$$\Omega_n = \{ \hat{\sigma}^2 > (1-2\delta)\sigma^2 \}.$$

On Ω_n , we know that

$$\forall m \in \mathcal{M}, \text{pen}(m) \geq (1+2\eta) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2.$$

Taking care of the random nature of the penalty, we argue as in the proof of Theorem 3.2 with $L_m = \eta$ to get

$$\begin{aligned}
\mathbb{E} [\|s - \tilde{s}\|_n^2 \mathbb{1}_{\Omega_n}] &\leq \frac{\eta+1}{\eta} \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \mathbb{E}[\text{pen}(m)] - \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \right\} \\
&\quad + \frac{\rho^2(P_n) \sigma^2}{n} R''_{P_n, \eta}(\mathcal{F})
\end{aligned} \tag{3.7.27}$$

where $R''_{P_n, \eta}(\mathcal{F})$ is defined by

$$R''_{P_n, \eta}(\mathcal{F}) = \frac{2(1+\eta)^4}{\eta^3} \sum_{m \in \mathcal{M}} \exp \left(-\frac{\eta^3}{2(1+\eta)^3} \times \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} \right).$$

We use Lemma 3.11 and (3.5.1) to get an upperbound for $\mathbb{E}[\text{pen}(m)]$,

$$\begin{aligned}\mathbb{E}[\text{pen}(m)] &\leq (1+\theta) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 + (1+\theta) \frac{\text{Tr}(^t P_n \pi_m P_n) \|s - \pi s\|_n^2}{\text{Tr}(^t P_n (I_n - \pi) P_n)} \\ &\leq (1+\theta) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 + (1+\theta) \frac{\text{Tr}(^t P_n P_n) \|s - \pi s\|_n^2}{\text{Tr}(^t P_n (I_n - \pi) P_n)} \\ &\leq (1+\theta) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 + 2(1+\theta) \|s - \pi s\|_n^2.\end{aligned}$$

The Proposition 3.1 and (3.7.27) give

$$\begin{aligned}\mathbb{E}[\|s - \tilde{s}\|_n^2 \mathbb{1}_{\Omega_n}] &\leq C(\theta) \inf_{m \in \mathcal{M}} \mathbb{E}[\|s - \hat{s}_m\|_n^2] + 2(\theta+1) \|s - \pi s\|_n^2 \\ &\quad + \frac{\rho^2(P_n) \sigma^2}{n} R''_{P_n, \eta}(\mathcal{F})\end{aligned}\tag{3.7.28}$$

where $C(\theta) > 1$.

We now bound $\mathbb{E}[\|s - \tilde{s}\|_n^2 \mathbb{1}_{\Omega_n^c}]$. Note that

$$\begin{aligned}\|s - \tilde{s}\|_n^2 &= \|s - s_{\hat{m}}\|_n^2 + \sigma^2 \|\pi_{\hat{m}} P_n \varepsilon\|_n^2 \\ &\leq \|s\|_n^2 + \sigma^2 \|P_n \varepsilon\|_n^2\end{aligned}$$

and thus, by the Cauchy–Schwarz Inequality,

$$\begin{aligned}\mathbb{E}[\|s - \tilde{s}\|_n^2 \mathbb{1}_{\Omega_n^c}] &\leq \|s\|_n^2 \mathbb{P}(\Omega_n^c) + \sigma^2 \mathbb{E}[\|P_n \varepsilon\|_n^2 \mathbb{1}_{\Omega_n^c}] \\ &\leq \left(\|s\|_n^2 + \sigma^2 \mathbb{E}[\|P_n \varepsilon\|_n^4]^{1/2} \right) \mathbb{P}(\Omega_n^c)^{1/2}.\end{aligned}$$

Moreover, the eigenvalues of the matrix $P_n {}^t P_n$ are nonnegative and so

$$\begin{aligned}\mathbb{E}[\|P_n \varepsilon\|_n^4]^{1/2} &= \left(\text{Var}(\|P_n \varepsilon\|_n^2) + \mathbb{E}[\|P_n \varepsilon\|_n^2]^2 \right)^{1/2} \\ &\leq \frac{1}{n} \sqrt{\text{Tr}(^t P_n P_n) (\text{Tr}(^t P_n P_n) + 2\rho^2(P_n))} \\ &\leq \frac{\text{Tr}(^t P_n P_n) + (\text{Tr}(^t P_n P_n) + 2\rho^2(P_n))}{2n} \\ &\leq \frac{\text{Tr}(^t P_n P_n) + \rho^2(P_n)}{n}.\end{aligned}$$

Finally, the Lemma 3.12 gives

$$\begin{aligned}\mathbb{E}[\|s - \tilde{s}\|_n^2 \mathbb{1}_{\Omega_n^c}] &\leq C'(\theta) \left(\|s\|_n^2 + \frac{\text{Tr}(^t P_n P_n) + \rho^2(P_n)}{n} \sigma^2 \right) \exp \left(-\frac{\theta^2 \text{Tr}(^t P_n P_n)}{32\rho^2(P_n)} \right) \\ &\leq C'(\theta) \left(\|s\|_n^2 + \frac{\rho^2(P_n)(n+1)}{n} \sigma^2 \right) \exp \left(-\frac{\theta^2 \text{Tr}(^t P_n P_n)}{32\rho^2(P_n)} \right) \\ &\leq C'(\theta) (\|s\|_n^2 + 2\rho^2(P_n) \sigma^2) \exp \left(-\frac{\theta^2 \text{Tr}(^t P_n P_n)}{32\rho^2(P_n)} \right)\end{aligned}\tag{3.7.29}$$

where $C'(\theta) > 1$. The inequality (3.5.4) follows by collecting (3.7.28) and (3.7.29).

Proof of Theorem 3.8. Given $\theta > 0$, we can find two positive numbers $\delta = \delta(\theta) < 1/3$ and $\eta = \eta(\theta)$ such that $(1+\theta)(1-3\delta) \geq (1+2\eta)$. Thus we define

$$\Omega'_n = \{ \hat{\sigma}^2 > (1-3\delta)\sigma^2 \}.$$

On Ω'_n , we know that

$$\forall m \in \mathcal{M}, \text{pen}(m) \geq (1+2\eta) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2.$$

Let \bar{m} be any element of \mathcal{M} that minimize $\|s - s_{m'}\|_n^2 + \sigma^2 \text{Tr}(^t P_n \pi_{m'} P_n)/n$ among $m' \in \mathcal{M}$. Taking care of the random nature of the penalty, we argue as in the proof of Theorem 3.3 with $L_m = \eta$ to get

$$\begin{aligned} \mathbb{E} [\|s - \tilde{s}\|_n^{2q} \mathbb{1}_{\Omega'_n}]^{1/q} &\leq C(q, \theta) \mathbb{E} \left[\left(\|s - s_{\bar{m}}\|_n^2 + \frac{\text{Tr}(^t P_n \pi_{\bar{m}} P_n)}{n} \hat{\sigma}^2 \right)^q \right]^{1/q} \\ &\quad + \frac{\rho^2(P_n) \sigma^2}{n} R_n(p, q, \theta)^{1/q} \end{aligned}$$

where

$$R_n(p, q, \theta) = C'(p, q, \theta) \tau_p \left[N_0 + \sum_{m \in \mathcal{M}: S \neq \{0\}} \left(1 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho(^t P_n \pi_m P_n)} \right) \left(\frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} \right)^{q-p/2} \right].$$

Since $q \leq 1$, by a convexity argument and Jensen's inequality we deduce

$$\begin{aligned} \mathbb{E} [\|s - \tilde{s}\|_n^{2q} \mathbb{1}_{\Omega'_n}]^{1/q} &\leq C(q, \theta) \left(\|s - s_{\bar{m}}\|_n^2 + \frac{\text{Tr}(^t P_n \pi_{\bar{m}} P_n)}{n} \mathbb{E}[\hat{\sigma}^2] \right) \\ &\quad + \frac{\rho^2(P_n) \sigma^2}{n} R_n(p, q, \theta)^{1/q}. \end{aligned} \quad (3.7.30)$$

Lemma 3.11 and (3.5.1) give

$$\begin{aligned} \frac{\text{Tr}(^t P_n \pi_{\bar{m}} P_n)}{n} \mathbb{E}[\hat{\sigma}^2] &= \frac{\text{Tr}(^t P_n \pi_{\bar{m}} P_n)}{n} \sigma^2 + \frac{n \text{Tr}(^t P_n \pi_{\bar{m}} P_n) \|s - \pi s\|_n^2}{n \text{Tr}(^t P_n (I_n - \pi) P_n)} \\ &\leq \frac{\text{Tr}(^t P_n \pi_{\bar{m}} P_n)}{n} \sigma^2 + 2 \|s - \pi s\|_n^2. \end{aligned}$$

Thus, by the definition of \bar{m} and Proposition 3.1, (3.7.30) becomes

$$\begin{aligned} \mathbb{E} [\|s - \tilde{s}\|_n^{2q} \mathbb{1}_{\Omega'_n}]^{1/q} &\leq C(q, \theta) \left(\inf_{m \in \mathcal{M}} \mathbb{E} [\|s - \hat{s}_m\|_n^2] + 2 \|s - \pi s\|_n^2 \right) \\ &\quad + \frac{\rho^2(P_n) \sigma^2}{n} R_n(p, q, \theta)^{1/q}. \end{aligned} \quad (3.7.31)$$

We now bound $\mathbb{E} [\|s - \tilde{s}\|_n^{2q} \mathbb{1}_{\Omega'^c_n}]$. Note that

$$\begin{aligned} \|s - \tilde{s}\|_n^2 &= \|s - s_{\bar{m}}\|_n^2 + \sigma^2 \|\pi_{\bar{m}} P_n \varepsilon\|_n^2 \\ &\leq \|s\|_n^2 + \sigma^2 \|P_n \varepsilon\|_n^2. \end{aligned}$$

Since $q \leq 1$, we have

$$\mathbb{E} [\|s - \tilde{s}\|_n^{2q} \mathbb{1}_{\Omega'^c_n}] \leq \|s\|_n^{2q} \mathbb{P}(\Omega'^c_n) + \sigma^{2q} \mathbb{E} [\|P_n \varepsilon\|_n^{2q} \mathbb{1}_{\Omega'^c_n}].$$

Hölder's inequality with exponent $p/2q > 1$ gives

$$\mathbb{E} [\|P_n \varepsilon\|_n^{2q} \mathbb{1}_{\Omega'^c_n}] \leq \mathbb{E} [\|P_n \varepsilon\|_n^p]^{2q/p} \mathbb{P}(\Omega'^c_n)^{1-2q/p}$$

and, since

$$\mathbb{E} [\|P_n \varepsilon\|_n^p]^{2q/p} \leq \rho^{2q}(P_n) \mathbb{E} [\|\varepsilon\|_n^p]^{2q/p} \leq \rho^{2q}(P_n) \tau_p^{2q/p},$$

we obtain by using Lemma 3.13 that

$$\begin{aligned} &\mathbb{E} [\|s - \tilde{s}\|_n^{2q} \mathbb{1}_{\Omega'^c_n}] \\ &\leq (\|s\|_n^{2q} + \sigma^{2q} \rho^{2q}(P_n) \tau_p^{2q/p}) \mathbb{P}(\Omega'^c_n)^{1-2q/p} \\ &\leq C(p, q, \theta) \kappa'_n(p, q, \theta) (\|s\|_n^{2q} + \sigma^{2q} \rho^{2q}(P_n) \tau_p^{2q/p}) (\tau_p \rho^{\alpha_p}(P_n) \text{Tr}(^t P_n P_n)^{-\beta_p})^{1-2q/p} \end{aligned}$$

where

$$\alpha_p = (p/2 - 1) \vee 1 \text{ and } \beta_p = (p/2 - 1) \wedge 1.$$

Thus, we get

$$\begin{aligned} & \mathbb{E}[\|s - \tilde{s}\|_n^{2q} \mathbb{1}_{\Omega_c^c}]^{1/q} \\ & \leq C'(p, q, \theta) \kappa_n(p, q, \theta) \tau_p^{1/q} (\|s\|_n^2 + \tau_p \rho^2(P_n) \sigma^2) \left(\frac{\rho^{2\alpha_p}(P_n)}{\text{Tr}(^t P_n P_n)^{\beta_p}} \right)^{1/q-2/p}. \end{aligned} \quad (3.7.32)$$

The announced result follows from (3.7.31) and (3.7.32).

3.7.3. Proofs of Corollaries and Propositions.

Proof of Corollary 3.4. Let us begin by applying Theorem 3.2 with constant weights $L_m = L$,

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq (1 + \theta^{-1}) \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + (\theta + L) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \right\} + \frac{\rho^2(P_n) \sigma^2}{n} R_n(\theta). \quad (3.7.33)$$

We now upperbound the remainder term. Assumption (\mathbf{A}'_3) and bounds on N_d and L lead to

$$\begin{aligned} R_n(\theta) & \leq \frac{2(1+\theta)^4}{\theta^3} \sum_{m \in \mathcal{M}} \exp \left(-\frac{\theta^2 L}{2(1+\theta)^3} \times \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} \right) \\ & \leq \frac{2(1+\theta)^4}{\theta^3} \sum_{m \in \mathcal{M}} \exp \left(-\frac{c\theta^2 L}{2(1+\theta)^3} \dim(S_m) \right) \\ & \leq \frac{2(1+\theta)^4}{\theta^3} \sum_{d \in \mathbb{N}} N_d e^{-(A+\omega)d} \\ & \leq \frac{2(1+\theta)^4}{\theta^3} \sum_{d \in \mathbb{N}} e^{-\omega d}. \end{aligned}$$

The last bound is clearly finite and we denote it by $R = R(\theta, \omega)$. Thus, we derive from (3.7.33)

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq (1 + \theta^{-1}) \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + ((\theta + L) \text{Tr}(^t P_n \pi_m P_n) + R \rho^2(P_n) (\dim(S_m) \vee 1)) \frac{\sigma^2}{n} \right\}$$

and hypothesis (\mathbf{A}'_3) gives

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq (1 + \theta^{-1}) \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + (\theta + L + R/c) (\text{Tr}(^t P_n \pi_m P_n) \vee c \rho^2(P_n)) \frac{\sigma^2}{n} \right\}$$

that concludes the proof.

Proof of Corollary 3.5. Since $p > 6$, we can take $q = 1$ and apply Theorem 3.3 with constant weights $L_m = L$ to get

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq C \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + (1 + \theta + L) \frac{\text{Tr}(^t P_n \pi_m P_n)}{n} \sigma^2 \right\} + \frac{\rho^2(P_n) \sigma^2}{n} R_n(p, 1, \theta). \quad (3.7.34)$$

To upperbound the remainder term, we use Assumption (\mathbf{A}'_3) and bounds on N_d and L to get

$$\begin{aligned} R_n(p, 1, \theta) & \leq C' \tau_p \left[1 + \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \left(1 + \frac{\text{Tr}(^t P_n \pi_m P_n)}{\rho(^t P_n \pi_m P_n)} \right) \left(\frac{L \text{Tr}(^t P_n \pi_m P_n)}{\rho^2(P_n)} \right)^{1-p/2} \right] \\ & \leq C' \tau_p \left[1 + \sum_{m \in \mathcal{M}: S_m \neq \{0\}} (1 + \dim(S_m)) (L c \dim(S_m))^{1-p/2} \right] \\ & \leq C' \tau_p \left[1 + \frac{(c\omega)^{1-p/2}}{A} \sum_{d>0} N_d (1+d) d^{1-p/2} \right] \\ & \leq C' \tau_p \left[1 + (c\omega)^{1-p/2} \sum_{d>0} (1+d)^{p/2-2-\omega} d^{1-p/2} \right]. \end{aligned}$$

The last bound is clearly finite and we denote it by $R\tau_p = R(\theta, p, \omega, \omega', c)\tau_p$. Thus, as we did in the previous proof, we derive from (3.7.34) and (\mathbf{A}'_3)

$$\mathbb{E} [\|s - \tilde{s}\|_n^2] \leq C'' \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + (1 + \theta + L + R\tau_p/c) (\text{Tr}({}^tP_n \pi_m P_n) \vee c\rho^2(P_n)) \frac{\sigma^2}{n} \right\}.$$

Since $\tau_p \geq 1$, the announced result follows.

Proof of Proposition 3.6. The collection \mathcal{F}^{DP} is nested and, for any $d \in \mathbb{N}$, we have $N_d \leq 1$. For Gaussian errors, Condition (3.2.6) is satisfied with $A = 0$ and, under moment condition, (3.2.8) is fulfilled with $A = 1$. In both cases, we are free to take $L = \theta = \eta/2$ and (\mathbf{A}_1) is true for $K = \eta$. Assumption (\mathbf{A}'_3) is fulfilled with $c = 1/\rho^2$ and, since $\dim(S_m) > 0$ for any $m \in \mathcal{M}$, we can apply Corollary 3.4 or 3.5 according to, respectively, $(\mathbf{H}_{\text{Gau\ss}})$ or $(\mathbf{H}_{\text{Mom}})$ holds. We argue in the same way than in Section 3.3 and we use (\mathbf{A}_3) to get

$$\begin{aligned} E [\|s - \tilde{s}\|_n^2] &\leq C \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{\text{Tr}({}^tP_n \pi_m P_n)}{n} \sigma^2 \right\} + C'(1 + \rho)^2 \left(\|t - \pi_{F+G}t\|_n^2 + \frac{R}{n} \sigma^2 \right) \\ &\leq C \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{\dim(S_m)}{n} \rho^2 \sigma^2 \right\} + C'(1 + \rho)^2 \left(\|t - \pi_{F+G}t\|_n^2 + \frac{R}{n} \sigma^2 \right). \end{aligned}$$

Thanks to the approximation properties of S_m and $F + G$, the following inequalities hold for any $s, t \in \mathcal{H}_\alpha(L)$ (see [DL93]),

$$\|s - s_m\|_n^2 \leq C(\alpha, L) \dim(S_m)^{-2\alpha}$$

and, since $F \perp G = (E + F)^\perp$,

$$\begin{aligned} \|t - \pi_{F+G}t\|_n^2 &= \|t\|_n^2 - \|\pi_F t\|_n^2 - \|\pi_G t\|_n^2 \\ &= \|t - \pi_F t\|_n^2 - \|t - \pi_{E+F}t\|_n^2 \\ &\leq \|t - \pi_F t\|_n^2 \\ &\leq C(\alpha, L) D_n^{-2\alpha}. \end{aligned}$$

Consequently, we obtain

$$E [\|s - \tilde{s}\|_n^2] \leq C'' \left(\inf_{m \in \mathcal{M}} \left\{ \dim(S_m)^{-2\alpha} + \frac{\dim(S_m)}{n} \right\} + D_n^{-2\alpha} + \frac{1}{n} \right).$$

Since $\alpha > \zeta_n$, we can consider some model in \mathcal{F}^{DP} with dimension D_m of order $n^{1/(2\alpha+1)}$ and derive that

$$E [\|s - \tilde{s}\|_n^2] \leq C'' \left(2n^{-2\alpha/(2\alpha+1)} + D_n^{-2\alpha} + \frac{1}{n} \right) \leq C_\alpha n^{-2\alpha/(2\alpha+1)}.$$

3.8. Lemmas

This section is devoted to some technical results and their proofs.

LEMMA 3.9. *Let $p, q > 0$ be two real numbers such that $2q < p$. For any $\theta > 0$, the following inequality holds*

$$\int_0^\infty \frac{qz^{q-1}}{(\theta + z)^{p/2}} dz \leq C(p, q) \theta^{q-p/2}$$

where $C(p, q) = p/(p - 2q)$.

PROOF. By splitting the integral around θ , we get

$$\begin{aligned} \int_0^\infty \frac{qz^{q-1}}{(\theta + z)^{p/2}} dz &= \int_0^\theta \frac{qz^{q-1}}{(\theta + z)^{p/2}} dz + \int_\theta^\infty \frac{qz^{q-1}}{(\theta + z)^{p/2}} dz \\ &\leq \theta^{-p/2} \int_0^\theta qz^{q-1} dz + \int_\theta^\infty qz^{q-1-p/2} dz \\ &\leq \left(1 + \frac{2q}{p - 2q} \right) \theta^{q-p/2}. \end{aligned}$$

□

The next lemma is a variant of a lemma due to Laurent and Massart.

LEMMA 3.10. *Let $A \in \mathbb{M}_n \setminus \{0\}$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ be a standard Gaussian vector of \mathbb{R}^n . For any $x > 0$, we have*

$$\mathbb{P} \left(n \|A\varepsilon\|_n^2 \geq \text{Tr}(A^t A) + 2\sqrt{\rho(A)^2 \text{Tr}(A^t A)x} + \rho(A)^2 x \right) \leq e^{-x} \quad (3.8.1)$$

and

$$\mathbb{P} \left(n \|A\varepsilon\|_n^2 \leq \text{Tr}(A^t A) - 2\sqrt{\rho(A)^2 \text{Tr}(A^t A)x} \right) \leq e^{-x} . \quad (3.8.2)$$

PROOF. It is known that $A\varepsilon$ is a centered Gaussian vector of \mathbb{R}^n of covariance matrix given by the positive symmetric matrix $A^t A$. Let us denote by $a_1, \dots, a_n \geq 0$ the eigenvalues of the $A^t A$. Thus, the distribution of $n \|A\varepsilon\|_n^2$ is the same as the one of $\sum_{i=1}^n a_i \varepsilon_i^2$. We have

$$\rho(A)^2 = \max_{i=1, \dots, n} |a_i| \quad \text{and} \quad \text{Tr}(A^t A) = \sum_{i=1}^n a_i .$$

Because the a_i 's are nonnegative,

$$\sum_{i=1}^n a_i^2 \leq \rho(A)^2 \text{Tr}(A^t A)$$

and we can apply the Lemma 1 of [LM00] to obtain the announced inequalities. □

We now introduce some properties that are satisfied by the estimator $\hat{\sigma}^2$ defined in (3.5.2).

LEMMA 3.11. *In the Gaussian case or under moment condition, the estimator $\hat{\sigma}^2$ satisfies*

$$\mathbb{E} [\hat{\sigma}^2] = \sigma^2 + \frac{n \|s - \pi s\|_n^2}{\text{Tr}(^t P_n (I_n - \pi) P_n)} .$$

PROOF. We have the following decomposition

$$\|Y - \pi Y\|_n^2 = \|s - \pi s\|_n^2 + \sigma^2 \|(I_n - \pi) P_n \varepsilon\|_n^2 + 2\sigma \langle s - \pi s, P_n \varepsilon \rangle_n . \quad (3.8.3)$$

The components of ε are independent and centered with unit variance. Thus, taking the expectation on both side, we obtain

$$\mathbb{E} [\|Y - \pi Y\|_n^2] = \|s - \pi s\|_n^2 + \sigma^2 \frac{\text{Tr}(^t P_n (I_n - \pi) P_n)}{n} .$$

□

LEMMA 3.12. *Consider the estimator $\hat{\sigma}^2$ defined in the Gaussian case. For any $0 < \delta < 1/2$,*

$$\mathbb{P} (\hat{\sigma}^2 \leq (1 - 2\delta)\sigma^2) \leq C_\delta \exp \left(-\frac{\delta^2 \text{Tr}(^t P_n P_n)}{16\rho^2(P_n)} \right)$$

where $C_\delta > 1$ can be taken equal to $1 + \exp(\delta/2)$.

PROOF. Let $a \in V^\perp$ such that $\|a\|_n^2 = 1$, we set

$$u = \begin{cases} (s - \pi s) / \|s - \pi s\|_n & \text{if } s \neq \pi s , \\ a & \text{otherwise} . \end{cases}$$

We have

$$\begin{aligned} 2\sigma |\langle s - \pi s, P_n \varepsilon \rangle_n| &= 2\sigma |\langle u, P_n \varepsilon \rangle_n| \times \|s - \pi s\|_n \\ &\leq \|s - \pi s\|_n^2 + \sigma^2 \langle u, P_n \varepsilon \rangle_n^2 \end{aligned}$$

and we deduce from (3.8.3)

$$\begin{aligned} \|Y - \pi Y\|_n^2 &\geq \sigma^2 \|(I_n - \pi) P_n \varepsilon\|_n^2 - \sigma^2 \langle u, P_n \varepsilon \rangle_n^2 \\ &= \sigma^2 (\|P_n \varepsilon\|_n^2 - (\|\pi P_n \varepsilon\|_n^2 + \langle u, P_n \varepsilon \rangle_n^2)) \\ &= \sigma^2 (\|P_n \varepsilon\|_n^2 - \|\pi' P_n \varepsilon\|_n^2) \end{aligned} \quad (3.8.4)$$

where π' is the orthogonal projection onto $V \oplus \mathbb{R}u$. Consequently,

$$\begin{aligned} \mathbb{P}(\hat{\sigma} \leq (1 - 2\delta)\sigma^2) &\leq \mathbb{P}(n\|P_n\varepsilon\|_n^2 - n\|\pi'P_n\varepsilon\|_n^2 \leq (1 - 2\delta)\text{Tr}({}^tP_n(I_n - \pi)P_n)) \\ &\leq \mathbb{P}(n\|P_n\varepsilon\|_n^2 - \text{Tr}({}^tP_nP_n) \leq -\delta\text{Tr}({}^tP_n(I_n - \pi)P_n)) \\ &\quad + \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \text{Tr}({}^tP_n\pi P_n) \geq \delta\text{Tr}({}^tP_n(I_n - \pi)P_n)) \\ &= \mathbb{P}_1 + \mathbb{P}_2. \end{aligned} \quad (3.8.5)$$

The Inequality (3.8.2) and (3.5.1) give us the following upperbound for \mathbb{P}_1 ,

$$\mathbb{P}_1 \leq \exp\left(-\frac{\delta^2\text{Tr}({}^tP_n(I_n - \pi)P_n)^2}{4\rho^2(P_n)\text{Tr}({}^tP_nP_n)}\right) \leq \exp\left(-\frac{\delta^2\text{Tr}({}^tP_nP_n)}{16\rho^2(P_n)}\right). \quad (3.8.6)$$

By the properties of the norm ρ , we deduce that

$$\text{Tr}({}^tP_n\pi'P_n) = \text{Tr}({}^tP_n\pi P_n) + \text{Tr}({}^tP_n\pi_u P_n) \leq \text{Tr}({}^tP_n\pi P_n) + \rho^2(P_n) \quad (3.8.7)$$

where we have defined π_u as the orthogonal projection onto $\mathbb{R}u$. We now apply (3.8.1) with $A = \pi'P_n$ to obtain, for any $x > 0$,

$$\begin{aligned} \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \text{Tr}({}^tP_n\pi P_n) \geq \delta\text{Tr}({}^tP_n\pi P_n)/2 + (1 + \delta/2)\rho^2(P_n) + (1 + 2/\delta)x) \\ \leq \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 \geq (1 + \delta/2)\text{Tr}({}^tP_n\pi'P_n) + (1 + 2/\delta)x) \\ \leq \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \text{Tr}({}^tP_n\pi'P_n) \geq 2\sqrt{\text{Tr}({}^tP_n\pi'P_n)x} + x) \\ \leq \exp(-x/\rho(\pi'P_n)^2) \\ \leq \exp(-x/\rho^2(P_n)). \end{aligned}$$

Obviously, this inequality can be extended to $x \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \text{Tr}({}^tP_n\pi P_n) \geq \delta\text{Tr}({}^tP_n\pi P_n)/2 + (1 + \delta/2)\rho^2(P_n) + (1 + 2/\delta)x) \\ \leq \exp\left(-\frac{x \vee 0}{\rho^2(P_n)}\right) \end{aligned} \quad (3.8.8)$$

and we take

$$\begin{aligned} x &= \frac{\delta}{\delta + 2} \left(\delta\text{Tr}({}^tP_n(I_n - \pi)P_n) - \frac{\delta}{2}\text{Tr}({}^tP_n\pi P_n) - \left(1 + \frac{\delta}{2}\right)\rho^2(P_n) \right) \\ &= \frac{\delta}{\delta + 2} \left(\delta\text{Tr}({}^tP_nP_n) - \frac{3\delta}{2}\text{Tr}({}^tP_n\pi P_n) - \left(1 + \frac{\delta}{2}\right)\rho^2(P_n) \right) \\ &\geq \frac{\delta}{\delta + 2} \left(\frac{\delta\text{Tr}({}^tP_nP_n)}{4} - \left(1 + \frac{\delta}{2}\right)\rho^2(P_n) \right). \end{aligned}$$

Finally, we get

$$\begin{aligned} \mathbb{P}_2 &\leq \exp\left(-\frac{\delta}{(\delta + 2)\rho^2(P_n)} \left(\frac{\delta\text{Tr}({}^tP_nP_n)}{4} - \left(1 + \frac{\delta}{2}\right)\rho^2(P_n) \right)_+\right) \\ &\leq \exp\left(-\frac{\delta}{2} \left(\frac{\delta\text{Tr}({}^tP_nP_n)}{2(\delta + 2)\rho^2(P_n)} - 1 \right)_+\right) \\ &= \left\{ e^{\delta/2} \exp\left(-\frac{\delta^2\text{Tr}({}^tP_nP_n)}{4(\delta + 2)\rho^2(P_n)}\right) \right\} \wedge 1. \end{aligned} \quad (3.8.9)$$

To conclude, we use (3.8.6) and (3.8.9) in (3.8.5). \square

LEMMA 3.13. *Consider the estimator $\hat{\sigma}^2$ defined under moment condition. For any $0 < \delta < 1/3$, there exists a sequence $(\kappa_{\delta,n})_{n \in \mathbb{N}}$ of positive numbers that tends to a positive constant κ_δ as $\text{Tr}({}^tP_nP_n)/\rho^2(P_n)$ tends to infinity, such that*

$$\mathbb{P}(\hat{\sigma}^2 \leq (1 - 3\delta)\sigma^2) \leq C(p, \delta)\kappa_{\delta,n}\tau_p\rho^{(p-2) \vee 2}(P_n)\text{Tr}({}^tP_nP_n)^{-((p/2-1) \wedge 1)}.$$

PROOF. We define the vector $u \in V^\perp$ and the projection matrix π' as we did in the proof of Lemma 3.12. The lowerbound (3.8.4) does not depend on the distribution of ε and gives

$$\begin{aligned} \mathbb{P}(\hat{\sigma}^2 \leq (1 - 3\delta)\sigma^2) \\ \leq \mathbb{P}(n\|P_n\varepsilon\|_n^2 - n\|\pi'P_n\varepsilon\|_n^2 \leq (1 - 3\delta)\text{Tr}(^tP_n(I_n - \pi)P_n)) . \end{aligned} \quad (3.8.10)$$

Since the matrix tP_nP_n is symmetric, we have the following decomposition

$$\begin{aligned} n\|P_n\varepsilon\|_n^2 - \text{Tr}(^tP_nP_n) &= n\langle ^tP_nP_n\varepsilon, \varepsilon \rangle_n - \text{Tr}(^tP_nP_n) \\ &= \sum_{i=1}^n \sum_{j=1}^n (^tP_nP_n)_{ij} \varepsilon_i \varepsilon_j - \text{Tr}(^tP_nP_n) \\ &= \sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1) + 2 \sum_{i=1}^n \sum_{j>i} (^tP_nP_n)_{ij} \varepsilon_i \varepsilon_j . \end{aligned}$$

Thus, (3.8.10) leads to

$$\mathbb{P}(\hat{\sigma}^2 \leq (1 - 3\delta)\sigma^2) \leq \bar{\mathbb{P}}_1 + \bar{\mathbb{P}}_2 + \bar{\mathbb{P}}_3 \quad (3.8.11)$$

where we have set

$$\begin{aligned} \bar{\mathbb{P}}_1 &= \mathbb{P}\left(\sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1) \leq -\delta \text{Tr}(^tP_n(I_n - \pi)P_n)\right) , \\ \bar{\mathbb{P}}_2 &= \mathbb{P}\left(2 \sum_{i=1}^n \sum_{j>i} (^tP_nP_n)_{ij} \varepsilon_i \varepsilon_j \leq -\delta \text{Tr}(^tP_n(I_n - \pi)P_n)\right) \end{aligned}$$

and

$$\bar{\mathbb{P}}_3 = \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \text{Tr}(^tP_n\pi P_n) \geq \delta \text{Tr}(^tP_n(I_n - \pi)P_n)) .$$

Note that $\bar{\mathbb{P}}_1$ concerns a sum of independent centered random variables. By Markov's inequality and (3.5.1), we get

$$\begin{aligned} \bar{\mathbb{P}}_1 &\leq \mathbb{P}\left(\left|\sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1)\right| \geq \delta \text{Tr}(^tP_n(I_n - \pi)P_n)\right) \\ &\leq \delta^{-p/2} \text{Tr}(^tP_n(I_n - \pi)P_n)^{-p/2} \mathbb{E}\left[\left|\sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1)\right|^{p/2}\right] \\ &\leq 2^{p/2} \delta^{-p/2} \text{Tr}(^tP_nP_n)^{-p/2} \mathbb{E}\left[\left|\sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1)\right|^{p/2}\right] . \end{aligned} \quad (3.8.12)$$

If $p \geq 4$ then we use the Rosenthal Inequality (see Chapter 2 of [Pet95]) and (3.7.23) to obtain

$$\mathbb{E}\left[\left|\sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1)\right|^{p/2}\right] \leq C'(p)\tau_p \left(\sum_{i=1}^n (^tP_nP_n)_{ii}^{p/2} + \left(\sum_{i=1}^n (^tP_nP_n)_{ii}^2\right)^{p/4}\right) .$$

Since, for any $i \in \{1, \dots, n\}$, $(^tP_nP_n)_{ii} \leq \rho^2(P_n)$, by a convexity argument, we get

$$\mathbb{E}\left[\left|\sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1)\right|^{p/2}\right] \leq 2C'(p)\tau_p \rho^{p/2}(P_n) \text{Tr}(^tP_nP_n)^{p/4} .$$

If $2 < p < 4$, we refer to [vBE65] for the following inequality

$$\begin{aligned} \mathbb{E}\left[\left|\sum_{i=1}^n (^tP_nP_n)_{ii} (\varepsilon_i^2 - 1)\right|^{p/2}\right] &\leq 2 \sum_{i=1}^n |(^tP_nP_n)_{ii} (\varepsilon_i^2 - 1)|^{p/2} \\ &\leq C''(p)\tau_p \rho^{p-2}(P_n) \text{Tr}(^tP_nP_n) . \end{aligned}$$

In both cases, (3.8.12) becomes

$$\bar{\mathbb{P}}_1 \leq C(p)\delta^{-p/2}\tau_p\rho^{p/2}(P_n)\mathrm{Tr}({}^tP_nP_n)^{-\beta} \quad (3.8.13)$$

with $\beta = (p/2 - 1) \wedge p/4$.

Let us now bound $\bar{\mathbb{P}}_2$. By Chebyshev's inequality, we get

$$\begin{aligned} \bar{\mathbb{P}}_2 &\leq \mathbb{P}\left(\left|2\sum_{i=1}^n\sum_{j>i}({}^tP_nP_n)_{ij}\varepsilon_i\varepsilon_j\right| \geq \delta\mathrm{Tr}({}^tP_n(I_n - \pi)P_n)\right) \\ &\leq \delta^{-2}\mathrm{Tr}({}^tP_n(I_n - \pi)P_n)^{-2}\mathbb{E}\left[\left(2\sum_{i=1}^n\sum_{j>i}({}^tP_nP_n)_{ij}\varepsilon_i\varepsilon_j\right)^2\right] \\ &\leq 4\delta^{-2}\mathrm{Tr}({}^tP_nP_n)^{-2}\sum_{i=1}^n\sum_{j>i}^n\sum_{p=1}^n\sum_{q>p}({}^tP_nP_n)_{ij}({}^tP_nP_n)_{pq}\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_p\varepsilon_q]. \end{aligned}$$

Note that, by independence between the components of ε , the expectation in the last sum is nonnull if and only if $i = p$ and $j = q$ (in this case, its value is 1). Thus, we have

$$\begin{aligned} \bar{\mathbb{P}}_2 &\leq 4\delta^{-2}\mathrm{Tr}({}^tP_nP_n)^{-2}\sum_{i=1}^n\sum_{j>i}({}^tP_nP_n)_{ij}^2 \\ &\leq 4\delta^{-2}\mathrm{Tr}({}^tP_nP_n)^{-2}\mathrm{Tr}(({}^tP_nP_n)^2) \\ &\leq 4\delta^{-2}\rho^2(P_n)\mathrm{Tr}({}^tP_nP_n)^{-1}. \end{aligned} \quad (3.8.14)$$

We finally focus on $\bar{\mathbb{P}}_3$. Recalling (3.8.7), we apply Corollary 5.1 of [Bar00] with $\tilde{A} = {}^tP_n\pi'P_n$ to obtain, for any $x > 0$,

$$\begin{aligned} \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_n\pi P_n) \geq \delta\mathrm{Tr}({}^tP_n\pi P_n)/2 + (1 + \delta/2)\rho^2(P_n) + (1 + 2/\delta)x) \\ \leq \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 \geq (1 + \delta/2)\mathrm{Tr}({}^tP_n\pi'P_n) + (1 + 2/\delta)x) \\ \leq \mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_n\pi'P_n) \geq 2\sqrt{\mathrm{Tr}({}^tP_n\pi'P_n)}x + x) \\ \leq C(p)\tau_p\mathrm{Tr}({}^tP_n\pi'P_n)\rho(\pi'P_n)^{p-2}x^{-p/2} \\ \leq C(p)\tau_p\mathrm{Tr}({}^tP_nP_n)\rho^{p-2}(P_n)x^{-p/2}. \end{aligned}$$

Thus, for any $x \in \mathbb{R}$, we define

$$\psi(x) = \begin{cases} C(p)\tau_p\mathrm{Tr}({}^tP_nP_n)\rho^{p-2}(P_n)x^{-p/2} \wedge 1 & \text{if } x > 0 \\ 1 & \text{if } x \leq 0 \end{cases}$$

and $\psi(x)$ is an upperbound for

$$\mathbb{P}(n\|\pi'P_n\varepsilon\|_n^2 - \mathrm{Tr}({}^tP_n\pi P_n) \geq \delta\mathrm{Tr}({}^tP_n\pi P_n)/2 + (1 + \delta/2)\rho^2(P_n) + (1 + 2/\delta)x).$$

If we take

$$\begin{aligned} x &= \frac{\delta}{\delta+2}\left(\delta\mathrm{Tr}({}^tP_n(I_n - \pi)P_n) - \frac{\delta}{2}\mathrm{Tr}({}^tP_n\pi P_n) - \left(1 + \frac{\delta}{2}\right)\rho^2(P_n)\right) \\ &= \frac{\delta}{\delta+2}\left(\delta\mathrm{Tr}({}^tP_nP_n) - \frac{3\delta}{2}\mathrm{Tr}({}^tP_n\pi P_n) - \left(1 + \frac{\delta}{2}\right)\rho^2(P_n)\right) \\ &\geq \frac{\delta}{\delta+2}\left(\frac{\delta\mathrm{Tr}({}^tP_nP_n)}{4} - \left(1 + \frac{\delta}{2}\right)\rho^2(P_n)\right) \end{aligned}$$

then we obtain

$$\begin{aligned}
\bar{\mathbb{P}}_3 &\leq C'(p, \delta) \tau_p \frac{\text{Tr}({}^t P_n P_n) \rho^{p-2}(P_n)}{(\delta \text{Tr}({}^t P_n P_n)/4 - (1 + \delta/2) \rho^2(P_n))_+^{p/2}} \wedge 1 \\
&\leq C''(p, \delta) \tau_p \frac{\text{Tr}({}^t P_n P_n)^{1-p/2} \rho^{p-2}(P_n)}{(1 - 2(1 + 2/\delta) \rho^2(P_n)/\text{Tr}({}^t P_n P_n))_+^{p/2}} \wedge 1
\end{aligned} \tag{3.8.15}$$

To conclude, we use (3.8.13), (3.8.14) and (3.8.15) in 3.8.11. \square

Quadratic risk of the LSE and the MLE in a Gaussian framework with dependent data : an example

In this appendix, we consider the framework (3.1.7) given by the n dimensional vector

$$Y = s + \sigma P\varepsilon$$

where $s = (s_1, \dots, s_n)' \in \mathbb{R}^n$ and $\sigma > 0$ are unknown, P is some known square matrix of size n and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ is an unobservable standard Gaussian vector. The estimation of the vector s in such a framework for a general matrix P was one of the objects of Chapter 3. In the introduction of this chapter, we mentioned some differences between the risks of the LSE (Least-Squares Estimator) and the MLE (Maximum Likelihood Estimator) of s in the case of an invertible matrix P . This appendix is devoted to illustrate these different behaviours on a particular example.

A.1. Notations and recalls

Hereafter, we mainly use the same notations than in Chapter 3. Let us introduce some known square matrix $A = (a_{ij})_{i,j=1,2}$. We assume that A is invertible and we denote by $A^{-1} = (\alpha_{ij})_{i,j=1,2}$ its inverse. Thus, we define the square matrix P of size $2n$ by the matrix with n times A on the diagonal, namely

$$P = \begin{pmatrix} A & \mathbf{0}_{2,2} & \cdots & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \cdots & \mathbf{0}_{2,2} & A \end{pmatrix}$$

where $\mathbf{0}_{k,k'}$ stands for the null matrix of size $k \times k'$. The invertibility of A implies that of P and we have

$$P^{-1} = \begin{pmatrix} A^{-1} & \mathbf{0}_{2,2} & \cdots & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0}_{2,2} \\ \mathbf{0}_{2,2} & \cdots & \mathbf{0}_{2,2} & A^{-1} \end{pmatrix}.$$

For any $m \in \mathcal{M} = \{0, \dots, n\}$, we consider the model S_m generated by the m first coordinates,

$$S_m = \{(x_1, \dots, x_m, 0, \dots, 0)' \in \mathbb{R}^n : x_1, \dots, x_m \in \mathbb{R}\}.$$

Clearly, the dimension of S_m is equal to m and we denote by π_m the orthogonal projector onto S_m . The LSE of s in S_m is the projection of the data, $\hat{s}_m = \pi_m Y$. Proposition 3.1 gives its quadratic risk

$$\mathbb{E} [\|s - \hat{s}_m\|_n^2] = \|s - s_m\|_n^2 + \frac{\text{Tr}(^t P \pi_m P)}{n} \sigma^2 \quad (\text{A.1.1})$$

where $s_m = \pi_m s$. Since P is invertible, we also can consider the likelihood function according to Y . Denoting by $\pi_m^{(P)}$ the orthogonal projection onto $P^{-1}S_m$, we derive from an easy computation that

the MLE of s in S_m is equal to

$$\hat{s}_m^L = P\pi_m^{(P)}P^{-1}Y.$$

The identity (3.1.11) corresponds to its quadratic risk,

$$\mathbb{E} \left[\|s - \hat{s}_m^L\|_n^2 \right] = \left\| s - P\pi_m^{(P)}P^{-1}s \right\|_n^2 + \frac{\text{Tr}(P\pi_m^{(P)}{}^tP)}{n} \sigma^2. \quad (\text{A.1.2})$$

A.2. Study of the risk

In both cases, the order of the risk is a sum of two terms. The first one is called the bias term and corresponds to the capacity to approximate the true value of s . The other one, called variance term, is similar to a dimensional term and measures the complexity of the underlying model. We saw in Chapter 3 that we need to find some trade-off between these two quantities in order to get a small quadratic risk. Note that, since $P\pi_m^{(P)}P^{-1}s \in S_m$, the bias in (A.1.1) is always no larger than the bias in (A.1.2). We also claimed that the variances terms in (A.1.1) and (A.1.2) are similar. Indeed, in our example, if $0 < \rho_1 \leq \rho_2$ are the two eigenvalues of the symmetric matrix A^tA , then we know

$$\rho_1 m \leq \text{Tr}({}^tP\pi_m P) \leq \rho_2 m \quad \text{and} \quad \rho_1 m \leq \text{Tr}({}^tP\pi_m^{(P)}P) \leq \rho_2 m.$$

Consequently, up to some known multiplicative factor, the variance terms are comparable. These remarks led us to restrict our study only to the LSE in Chapter 3. In order to clarify this comparison between the risks of the LSE and the MLE, we now explicitly compute them in our particular framework.

To this end, we introduce the vector

$$v = \begin{cases} \begin{pmatrix} 1 \\ \alpha_{21}/\alpha_{11} \end{pmatrix} & , \text{ if } \alpha_{11} \neq 0, \\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} & , \text{ otherwise.} \end{cases}$$

Let us first assume that the dimension of S_m is even, namely $m = 2k$. In such a case, note that $P^{-1}S_{2k} = S_{2k}$ and, thus, $\pi_{2k} = \pi_{2k}^{(P)}$. So, the variance terms are both equal to $k\text{Tr}(A^tA)\sigma^2/n$. Moreover, we easily obtain that the bias terms are the same. Consequently, for an even dimensional model, we have

$$\mathbb{E} [\|s - \hat{s}_{2k}\|_n^2] = \mathbb{E} [\|s - \hat{s}_{2k}^L\|_n^2].$$

We now focus on the odd case $m = 2k + 1$. The projection matrices can be expanded as follows in the canonical basis,

$$\pi_{2k+1} = \begin{pmatrix} I_{2k+1} & \mathbf{0}_{2k+1, n-2k-1} \\ \mathbf{0}_{n-2k-1, 2k+1} & \mathbf{0}_{n-2k-1, n-2k-1} \end{pmatrix} \quad \text{and} \quad \pi_{2k+1}^{(P)} = \begin{pmatrix} \tilde{I}_k & \mathbf{0}_{2k+2, n-2k-1} \\ \mathbf{0}_{n-2k-2, 2k+1} & \mathbf{0}_{n-2k-2, n-2k-1} \end{pmatrix}$$

where \tilde{I}_k is the matrix of size $(2k+2) \times (2k+1)$ given by

$$\tilde{I}_k = \begin{pmatrix} I_{2k} & \mathbf{0}_{2k, 1} \\ \mathbf{0}_{2, 2k} & v \end{pmatrix}.$$

Thus, we can compute the trace in the variance term of (A.1.1),

$$\begin{aligned} \text{Tr}({}^tP\pi_{2k+1}P) &= k\text{Tr}(A^tA) + (A^tA)_{11} \\ &= k\text{Tr}(A^tA) + a_{11}^2 + a_{12}^2. \end{aligned}$$

Note that the condition $\alpha_{11} = 0$ is equivalent to $a_{22} = 0$. So, by classical linear algebra results, we obtain the value of the trace in (A.1.2),

$$\begin{aligned} \text{Tr}(P\pi_{2k+1}^{(P)}{}^tP) &= k\text{Tr}(A^tA) + \frac{a_{11}}{\alpha_{11} + \alpha_{21}\mathbb{1}_{\alpha_{11}=0}} \\ &= k\text{Tr}(A^tA) + \frac{a_{11}(a_{11}a_{22} - a_{12}a_{21})}{a_{22} - a_{21}\mathbb{1}_{a_{22}=0}}. \end{aligned}$$

Since A is invertible, the denominator is not null. These identities illustrate the fact that the variance terms of the LSE and the MLE in an odd dimensional model are both of order of $k\text{Tr}(A^t A)\sigma^2/n$ up to some additional term. Consequently, according to the matrix A , we can compare the variance terms in the decomposition of the risks of \hat{s}_{2k+1} and \hat{s}_{2k+1}^L .

To conclude, we give some examples of matrix A that lead to various comparisons between the quadratic risks,

$$A_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \quad A_3 = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix}$$

and we denote by P_1 , P_2 and P_3 , respectively, the corresponding square matrices of size $2n$. For these matrices, we have

$$\text{Tr}(A_1^t A_1) = 3 \quad \text{Tr}(A_2^t A_2) = 7 \quad \text{Tr}(A_3^t A_3) = 7$$

and so, for $m = 2k + 1$,

$$\text{Tr}({}^t P_1 \pi_{2k+1} P_1) = 3k + 1 = \text{Tr}(P_1 \pi_{2k+1}^{(P_1)} {}^t P_1) ,$$

$$\text{Tr}({}^t P_2 \pi_{2k+1} P_2) = 7k + 5 > 7k + 2 = \text{Tr}(P_2 \pi_{2k+1}^{(P_2)} {}^t P_2) ,$$

and

$$\text{Tr}({}^t P_3 \pi_{2k+1} P_3) = 7k + 5 < 7k + 6 = \text{Tr}(P_3 \pi_{2k+1}^{(P_3)} {}^t P_3) .$$

For the choice A_2 , the MLE has a smaller variance term. Since the bias term is unknown, we can not compare the quadratic risks. For the matrices A_1 and A_3 , since we know that the bias term of the LSE is not larger than the one of the MLE, we deduce from the above inequalities that, for any $m \in \{0, \dots, n\}$,

$$\mathbb{E} [\|s - \hat{s}_m\|_n^2] \leq \mathbb{E} [\|s - \hat{s}_m^L\|_n^2] .$$

Bibliographie

- [Aka70] H. Akaike. Statistical predictor identification. *Annals of the Institute for Statistical Mathematics*, 1970.
- [Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In P.N. Petrov and F. Csaki, editors, *Proceedings 2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.
- [Aka74] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974.
- [Arl07] S. Arlot. *Rééchantillonnage et sélection de modèles*. PhD thesis, Université Paris 11, 2007.
- [Bar00] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.
- [Bar02] Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.
- [BBLM05] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Annals of Probability*, 33(2):514–560, 2005.
- [BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.
- [BC06] E. Brunel and F. Comte. Adaptive nonparametric regression estimation in presence of right censoring. *Mathematical Methods of Statistics*, 15(3):233–255, 2006.
- [BC08] E. Brunel and F. Comte. *Model selection for additive regression models in the presence of censoring*, chapter 1 in “Mathematical Methods in Survival Analysis, Reliability and Quality of Life”, pages 17–31. Wiley, 2008.
- [BCV01] Y. Baraud, F. Comte, and G. Viennet. Adaptive estimation in autoregression or β -mixing regression via model selection. *Annals of Statistics*, 29(3):839–875, 2001.
- [BF85] L. Breiman and J.H. Friedman. Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association*, 80(391):580–619, 1985.
- [BGH09] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Annals of Statistics*, 37(2):630–672, 2009.
- [BHT89] A. Buja, T.J. Hastie, and R.J. Tibshirani. Linear smoothers and additive models (with discussion). *Annals of Statistics*, 17:453–555, 1989.
- [BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. *Festschrift for Lucien Lecam: Research Papers in Probability and Statistics*, 1997.
- [BM01a] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

-
- [BM01b] L. Birgé and P. Massart. A generalized c_p criterion for gaussian model selection. *Prépublication 647, Universités de Paris 6 and Paris 7*, 2001.
 - [CR02] F. Comte and Y. Rozenholc. Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Processes and their Applications*, 97(1):111–145, 2002.
 - [CW08] T. Cai and L. Wang. Adaptive variance function estimation in heteroscedastic nonparametric regression. *Annals of Statistics*, 36(5):2025–2054, 2008.
 - [DJ94] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
 - [DL93] R.A. DeVore and G.G. Lorentz. *Constructive approximation*, volume 303. Springer-Verlag, 1993.
 - [Gau09] C.F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*. Perthes and Besser, Hamburg, 1809.
 - [Gen08] X. Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression. *Electron. J. Statist.*, 2:1345–1372, 2008.
 - [GP05] L. Galtchouk and S. Pergamenschikov. Efficient adaptive nonparametric estimation in heteroscedastic regression models. Université Louis Pasteur, IRMA, Preprint, 2005.
 - [GSL] GSL Development. *GSL: GNU Scientific Library*.
 - [HJ90] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.
 - [HMSW04] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, 2004.
 - [HT90] T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.
 - [Leg05] A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.
 - [Leo47] W. Leontief. Introduction to a theory of the internal structure of functional relationships. *Econometrica*, 15:361–373, 1947.
 - [Li87] K.C. Li. Asymptotic optimality for c_p , c_l , cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics*, 15:958–975, 1987.
 - [LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.
 - [LN95] O. Linton and J.P. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–101, 1995.
 - [Mal73] C.L. Mallows. Some comments on c_p . *Technometrics*, 15:661–675, 1973.
 - [Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6-23, 2003.
 - [MLN99] E. Mammen, O. Linton, and J.P. Nielsen. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27:1443–1490, 1999.
 - [MT98] A.D.R. McQuarrie and C.L. Tsai. *Regression and times series model selection*. River Edge, NJ, 1998.
 - [MvB09] L. Meier, S. van de Geer, and P. Bühlmann. *High-dimensional additive modeling*. To appear in *Annals of Statistics*, [arXiv:0806.4115](https://arxiv.org/abs/0806.4115), 2009.

-
- [OR97] J. Opsomer and D. Ruppert. Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25:186–211, 1997.
- [Pet95] V.V. Petrov. *Limit theorems of probability theory: sequences of independent random variables*. Oxford Studies in Probability 4, 1995.
- [PT90] B.T. Polyak and A.B. Tsybakov. Asymptotic optimality of the c_p - test for the orthogonal series estimation of regression. *Theory of probability and its applications*, 35:293–306, 1990.
- [R D07] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.
- [RLLW08] P.D. Ravikumar, H. Liu, J.D. Lafferty, and L.A. Wasserman. *SpAM: Sparse Additive Models*. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.
- [RW94] D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *Annals of Statistics*, 22(3):1346–1370, 1994.
- [Sch59] H. Scheffé. *The analysis of variance*. Wiley-Interscience, 1959.
- [Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.
- [Shi81] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.
- [SLS99] E. Severance-Lossin and S. Sperlich. Estimation of derivatives for additive separable models. *Statistics*, 33:241–265, 1999.
- [Sto85] C.J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 14(2):590–606, 1985.
- [TA94] D. Tjøstheim and B. Auestad. Nonparametric identification of nonlinear time series: Selecting significant lags. *Journal of the American Statistical Association*, 89:1410–1430, 1994.
- [vBE65] B. von Bahr and C.G. Esseen. Inequalities for the r th absolute moment of a sum of random variables $1 \leq r \leq 2$. *Annals of Mathematical Statistics*, 36:299–303, 1965.
- [Wt08] L. Wang and *al.* Effect of mean on variance function estimation in nonparametric regression. *Annals of Statistics*, 36(2):646–664, 2008.

Quand j'aurai mon niveau huit, j'achèterai une baliste,
Pour assiéger les donjons des nécromancultistes.
PEN OF CHAOS

Estimation par sélection de modèle en régression hétéroscédastique

Résumé : cette thèse s'inscrit dans les domaines de la statistique non-asymptotique et de la théorie statistique de la sélection de modèle. Son objet est la construction de procédures d'estimation de paramètres en régression hétéroscédastique. Ce cadre reçoit un intérêt croissant depuis plusieurs années dans de nombreux champs d'application. Les résultats présentés reposent principalement sur des inégalités de concentration et sont illustrés par des applications à des données simulées.

La première partie de cette thèse consiste dans l'étude du problème d'estimation de la moyenne et de la variance d'un vecteur gaussien à coordonnées indépendantes. Nous proposons une méthode de choix de modèle basée sur un critère de vraisemblance pénalisé. Nous validons théoriquement cette approche du point de vue non-asymptotique en prouvant des majorations de type oracle du risque de Kullback de nos estimateurs et des vitesses de convergence uniforme sur les boules de Hölder.

Un second problème que nous abordons est l'estimation de la fonction de régression dans un cadre hétéroscédastique à dépendances connues. Nous développons des procédures de sélection de modèle tant sous des hypothèses gaussiennes que sous des conditions de moment. Des inégalités oracles non-asymptotiques sont données pour nos estimateurs ainsi que des propriétés d'adaptativité. Nous appliquons en particulier ces résultats à l'estimation d'une composante dans un modèle de régression additif.

Mots-clés : statistique non-asymptotique, sélection de modèle, pénalisation, inégalité oracle, régression non-paramétrique, hétéroscédastique, modèle additif, adaptativité, vitesse minimax, risque de Kullback.

Model selection in heteroscedastic regression

Abstract : this thesis takes place within the theories of nonasymptotic statistics and model selection. Its goal is to provide data-driven procedures to estimate some parameters in heteroscedastic regression. This framework is receiving a large interest in various domains of applied mathematics. Our procedures rely in particular on some concentration inequalities and their practical efficiency is assessed on simulated data.

The first part is devoted to simultaneous estimation of the mean and the variance of a Gaussian vector with independent coordinates. To this end, we introduce a model selection procedure based on some penalized likelihood criterion. We prove nonasymptotic results for this method, such as oracle type inequalities and uniform convergence rates over Hölderian balls.

We also consider the problem of estimation of the regression function in an heteroscedastic regression framework with known dependencies. Model selection procedures are constructed for Gaussian errors and under moment conditions. Nonasymptotic oracle type inequalities and adaptivity are proved for the estimators. In particular, we apply these procedures to estimate a component in an additive regression model.

Keywords : nonasymptotic statistics, model selection, penalization, oracle inequality, nonparametric regression, heteroscedastic, additive model, adaptivity, minimax rate, Kullback risk.

AMS Classification : 62G08, 62J02